Electrostatic Interactions in Chemistry and Biology

Thesis by

Justin Patrick Gallivan

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

2000
(Defended May 22, 2000)

Acknowledgements

O.K., admit it. This is likely the only section of the thesis that you are going to read. Oh sure, you might flip through the rest. You might even pause to admire the fancy graphics on page 35, but the fact of the matter is, you really want to see if your name made it in here. Thus, I apologize in advance for anyone that I might inadvertently leave out.

First of all, I must thank my advisor, Dennis Dougherty. I don't think I could have asked for a better advisor. Dennis has provided a great environment in which to do science. He has given me the freedom to choose problems that I thought were interesting, and has been a constant source of advice and wisdom. But aside from being an excellent scientist, Dennis is quite simply a good person, and I have really enjoyed having the opportunity to work with him. Special thanks also go to Ellen Dougherty, who on many occasions welcomed the group into her home. Ellen realizes that beneath all of our academic bravado, we are just like elementary school kids who miss home.

I would also like to thank Henry Lester for allowing me to spend time in his lab. Although the draw of computational chemistry eventually led me back to the second floor of Crellin, I sincerely appreciate the time I spent on the third floor of Kerckhoff.

I also extend thanks to my committee. Peter Dervan has always offered sage advice. I will always remember a talk he gave about the importance of choosing good mentors – fortunately, I have had several. I would also like the thank Barbara Imperiali, Doug Rees, and John Bercaw, who have all taken time out of their busy schedules to attend committee meetings or write letters of recommendation on my behalf.

The Dougherty group has been a great place to work over the past several years. Kraig Anderson and Seth Miller made 247 Crellin a very interesting place during my first few years. Kraig's staunch cynicism and Seth's unwavering optimism nicely

balanced one another. Along with Kraig, Jon Forman and Mick Murray were always willing to go out for a couple of pints.[1] Sandro Mecozzi has to be one of the nicest guys in the world, and I appreciate having had the opportunity to work with him. Jeff Clites is one of the brightest people I have ever met and was always fun to talk with about any subject.

I owe special thanks to many people who helped me get started on various projects. Scott Silverman is a great teacher and friend. If they gave a Nobel Prize in knowledge of "The Simpsons," he would win it every year. I wish him all the best as he begins his academic career. Mark Nowak taught me most of the molecular biology that I know and I greatly appreciate that. Anthony West is an incredibly smart guy and he got me started down the computational chemistry path – this thesis would probably be a lot different without his influence. Wenge Zhong and I collaborated on the work described in Chapter 2, and the combination of theory and experiment made the project quite fun. Pam England and I had our differences at times, but I am glad that we got to interact for so many years, and I am a better person for it.

Of the "younger" students in the group, I would particularly like to thank Gabriel Brandt, Josh Maurer, and Don "Fancy Pants" Elmore (sorry, Don, it had to be done). Gabriel has always been willing to critically discuss almost any subject, scientific or otherwise, and he always offers an interesting opinion. Josh Maurer and Don Elmore are excellent people to bounce ideas off of, and reading the "Pint-Sized Punchlines" was always fun. 247 Crellin will be left in the very capable hands of Darren Beene, James Petersson, and Sarah Monahan. I have enjoyed having them around these last few months. I would also like to thank all of the other members of the Dougherty group that have made my time here special.

Outside of the group, I owe thanks to the Eastman Kodak Corporation for providing me with several years of funding, and to my three mentors there: Eric Ginsburg, Rakesh Jain, and Frank Coms. I learned a lot from all of you and always enjoyed my visits to Rochester. Special thanks also go to Dian Buchness, Tom Dunn, Steve

[1] It should be noted that along with the decline in the number of magnetism projects in the group came a decline in the amount of alcohol consumed by the group. Determining the significance of this is left as an exercise for the reader.

Gould, and Chris Smith. Aside from being very nice people, they have all gone above and beyond the call of duty to help me out in one way or another.

I would also like to thank Cory Nelson and Dave Lynn for being great friends. The trips to Las Vegas and San Francisco, poker nights, McMurphy's, and running the basketball pool were all lots of fun. I had a great time hanging out with these guys, and their presence at Caltech is missed. I would also like to thank Adam Matzger, who has been a good friend and an excellent source of wisdom. I wish he and Anna the best of luck as they start their professional careers.

Finally, I would like to thank my family. My wife, Martha, has been a constant source of support throughout my time here. Martha probably understands me better than I understand myself, and she always seems to know exactly the right thing to say. But most of all, Martha is always there to listen, and often that is the best thing. I would also like to thank my parents. They, too, have always been there for me and have supported me in everything that I have done. Mom, Dad, and Martha: I love you all.

Abstract

Electrostatic interactions such as hydrogen bonds, salt bridges, and cation-π interactions play a large role in structural biology. A major goal of this thesis is to build upon previous studies of the cation-π interaction to further understand its role in biological systems. Put simply, we wish to understand how, when, and why Nature uses cation-π interactions.

We begin by highlighting a cation-π interaction important in the binding of acetylcholine to the nicotinic acetylcholine receptor (nAChR). By combining ab initio calculations and molecular neurobiology, we provide compelling evidence that a cation-π interaction is a major determinant of the recognition of acetylcholine by the nAChR.

We then ask a broader question: To what extent does Nature use cation-π interactions within protein structures? By surveying the protein databank, we demonstrate that *energetically significant* cation-π interactions are quite common within protein structures. To explain why, we ask what advantages cation-π interactions have over other noncovalent interactions commonly found in proteins. Using quantum mechanical calculations, we study the strengths of cation-π interactions and salt bridges in both water and in a range of organic solvents. The results suggest that cation-π interactions maintain their strength over a wide range of solvents, whereas the strength of a salt bridge is severely attenuated when it is placed in a high-dielectric solvent.

We then turn our attention to a different type of electrostatic interaction – the interaction between water and hexafluorobenzene. We find that in the gas phase, water binds to hexafluorobenzene in a geometry in which the lone pairs of electrons located on the oxygen are directed towards the π-system of the aromatic. This surprising result is easily explained using electrostatics. In addition, we present computational studies of the triphenylene$\cdots$perfluorotriphenylene "supramolecular synthon."

Finally, we return to the nAChR. A challenge in the study of integral membrane proteins is determining their transmembrane topology. Here we present a potentially

general method for determining not only the transmembrane topology of a functional neuroreceptor expressed in a living cell, but also the surface accessibility of individual amino acids.

Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Introduction

One of the greatest achievements of modern organic chemistry is a thorough understanding of the nature of covalent bonding. Successes ranging from modern *ab initio* electronic structure methods to the total synthesis of complex molecules all testify to our ability to understand and manipulate covalent bonding. What is less well understood, however, is the nature of noncovalent bonding. It is the collective behavior of weak, noncovalent interactions that ultimately determines the structure and function of biological systems. As such, a thorough knowledge of weak interactions, such as hydrogen bonds, salt bridges, and the hydrophobic effect, is necessary to understand biological systems at the molecular level.

In addition to these well-studied interactions, the cation-π interaction, the electrostatically favorable interaction between a cation and a π-system, has been shown to be an important noncovalent binding interaction of relevance to structural biology (Figure 1.1) [1].

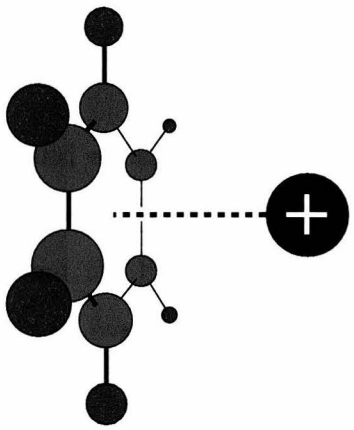
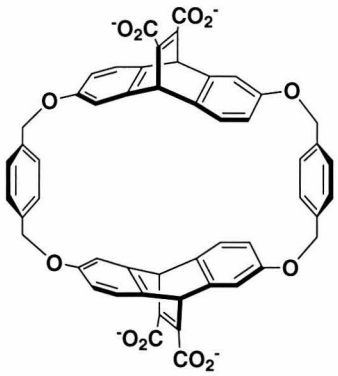


Figure 1.1: Schematic view of a cation binding to an aromatic – a cation-π interaction.

In the early 1980s, high-pressure mass spectrometry and ion cyclotron resonance studies demonstrated that cations bind to simple aromatics in the gas phase [2, 3]. In pioneering work, Kebarle showed that the interaction energy between K^+ and benzene in the gas phase (19 kcal/mol) is even stronger than the interaction between K^+ and water (18 kcal/mol) [2]. Later experiments have shown that in clusters containing both benzene and water molecules bound to a single cation, the benzenes are more

tightly bound than the water molecules [4]. The interaction is not limited to alkali metals, and it is now appreciated that organic cations such as tetramethylammonium and acetylcholine bind to aromatics in the gas phase [3, 5].

The cation-π interaction might have remained a gas phase curiosity were it not for work done by the Dougherty group in the 1980s and early 1990s. Using a series of water-soluble cyclophane hosts containing a hydrophobic cavity, e.g. **1**, Dougherty and co-workers demonstrated that these molecules could bind not only hydrophobic guests, but also cationic guests in aqueous media [6–11].



1

This is significant because to bind a cationic guest, a substantial (>40 kcal/mol) desolvation penalty must be paid to transfer the cation from bulk water to the interior of the host. This suggests that the binding energy gained in forming a cation-π interaction must be quite strong to overcome the desolvation penalty. A key result from the small molecule studies is that host molecules such as **1** bind the quaternary ammonium group of acetylcholine. This led to the prescient prediction by Dougherty and Stauffer that aromatic amino acids might be involved in the binding of acetylcholine in biological systems [9]. Indeed this is the case, and the x-ray structure of acetylcholinesterase reveals an "aromatic gorge" that presumably guides acetylcholine to the active site of the esterase [12]. Remarkably, the "anionic subsite" that contacts the quaternary ammonium group of acetylcholine is the aromatic sidechain of a tryptophan residue. This study clearly established a role for cation-π interactions in biological systems.

In addition to the host-guest studies, numerous computational studies have been undertaken to understand the physical basis of the cation-π interaction [13–17]. It is now appreciated that the cation-π interaction is a complex phenomenon, involving dispersion, polarization, and electrostatic interactions [1]. Interestingly, although the dispersion and polarization effects may represent a substantial contribution to the total binding energy in a cation-π interaction [14, 17], these effects remain relatively constant across a related series of aromatics [15]. As such, variations in the cation-π binding ability of aromatics correlate very well with variations in the electrostatic potential of the aromatic. Thus electrostatic potential maps of the aromatics, such as the one for benzene, shown in Figure 1.2, provide a useful guide to predicting the cation-π binding ability of a given aromatic [16].

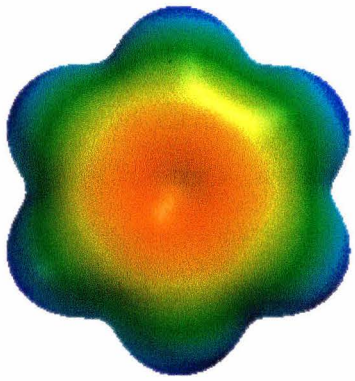


Figure 1.2: The calculated HF/6-31G** electrostatic potential surface of benzene. Red regions represent areas of negative electrostatic potential and blue regions represent positive electrostatic potential. *Scale: -25 to +25 kcal/mol.*

A major goal of this thesis is to build upon previous fundamental studies of the cation-π interaction to further understand the role of cation-π interactions in biological systems. Put simply, we wish to understand how, when, and perhaps more importantly, why Nature uses cation-π interactions. We do this in several different contexts, ranging from *ab initio* quantum mechanical calculations of the cation-π interaction both in the gas phase and in solution, to studies of the cation-π interaction between acetylcholine and the nicotinic acetylcholine receptor expressed in a living cell.

Chapter 2 highlights a particular cation-π interaction important in the binding

of acetylcholine to the nicotinic acetylcholine receptor. By combining high-level *ab initio* calculations and molecular neurobiology, we provide compelling evidence that a cation-π interaction is a major determinant of the recognition of acetylcholine by its natural receptor. In doing so, we provide some of the highest resolution structural information to date about the nicotinic acetylcholine receptor.

In Chapter 2, we only consider potential cation-π interactions between a particular receptor-ligand pair. The work presented in Chapter 3 addresses a broader question. In particular, we ask to what extent Nature uses cation-π interactions within protein structures. By surveying a large database of protein structures, we demonstrate that *energetically significant* cation-π interactions are quite common within protein structures. In fact, one quarter of all tryptophans in our dataset are involved in at least one cation-π interaction. Extensive statistics and the methodology used are presented.

The results in Chapter 3 present clear evidence that cation-π interactions are common in protein structures. In Chapter 4, we attempt to discover why this is the case. In particular, we ask what advantages cation-π interactions have over other non-covalent interactions commonly found in protein structures. Using high-level quantum mechanical calculations, we study the strengths of cation-π interactions and salt bridges in both water and in a range of organic solvents. The results suggest that cation-π interactions maintain their strength over a wide range of solvents, whereas the strength of a salt bridge is severely attenuated when it is placed in a high-dielectric solvent.

In Chapter 5, we turn our attention to a different type of electrostatic interaction. In particular, we computationally explore the interaction between two very simple molecules, water and hexafluorobenzene. We find that in the gas phase, water binds to hexafluorobenzene with 2.1 kcal/mol of binding energy. Most surprisingly, the geometry is such that that the lone pairs of electrons located on the oxygen are directed towards the π-system of the aromatic. This result would not normally be predicted from our views of orbital interactions, but it is easily explained using electrostatics. In addition to studies of the water$\cdots$hexafluorobenzene system, we present initial

computational studies of the triphenylene$\cdots$perfluorotriphenylene "supramolecular synthon."

In Chapter 6, we return our attention to the nicotinic acetylcholine receptor. A major challenge in the study of integral membrane proteins is determining the transmembrane topology of the protein. Here we present a potentially general method for determining not only the transmembrane topology of a functional neuroreceptor expressed in a living cell, but also the surface accessibility of individual amino acids.

Finally, in Chapter 7 we present experimental details of our efforts to develop new unnatural amino acids to explore membrane protein structure and function.

Bibliography

[1] J. C. Ma and D. A. Dougherty. The cation-π interaction. *Chem. Rev.*, 97:1303–1324, 1997.

[2] J. Sunner, K. Nishizawa, and P. Kebarle. Ion-solvent molecule interactions in the gas phase. The potassium ion and benzene. *J. Phys. Chem.*, 85:1814–1820, 1981.

[3] C.A. Deakyne and M. Meot-Ner (Mautner). Unconventional ionic hydrogen bonds. 2. NH^+ π. complexes of onium ions with olefins and benzene derivatives. *J. Am. Chem. Soc.*, 107:474–479, 1985.

[4] O. M. Cabarcos, C. J. Weinheimer, and J. M. Lisy. Size selectivity by cation-π interactions: Solvation of K^+ and Na^+ by benzene and water. *J. Chem. Phys.*, 110:8429–8435, 1999.

[5] C.A. Deakyne and M. Meot-Ner. Ionic hydrogen bonds in bioenergetics. 4. Interaction energies of acetylcholine with aromatic and polar molecules. *J. Am. Chem. Soc.*, 121(7):1546–1557, 1999.

[6] T.J. Shepodd, M.A. Petti, and D.A. Dougherty. Molecular recognition in aqueous media: Donor-acceptor and ion-dipole interactions produce tight binding for highly soluble guests. *J . Am. Chem. Soc.*, 110:1983–1985, 1988.

[7] M. A. Petti, T. J. Shepodd, R. E. Barrans Jr., and D. A. Dougherty. "Hydrophobic" binding of water-soluble guests by high-symmetry, chiral hosts. An electron-rich receptor site with a general affinity for quaternary ammonium compounds and electron-deficient π systems. *J. Am. Chem. Soc.*, 110:6825–6840, 1988.

[8] D.A. Stauffer, R.E. Barrans Jr., and D.A. Dougherty. Concerning the thermodynamics of molecular recognition in aqueous and organic media. Evidence for significant heat capacity effects. *J. Org. Chem.*, 55:2762–2767, 1990.

[9] D.A. Dougherty and D.A. Stauffer. Acetylcholine binding by a synthetic receptor. Implications for biological recognition. *Science*, 250:1558–1560, 1990.

[10] A. McCurdy, L. Jimenez, D. A. Stauffer, and D.A. Dougherty. Biomimetic catalysis of S_N2 reactions through cation-π interactions. The role of polarizability in catalysis. *J. Am. Chem. Soc.*, 114:10314–10321, 1992.

[11] P. C. Kearney, L. S. Mizoue, R. A. Kumpf, J. E. Forman, A. McCurdy, and D. A. Dougherty. Molecular recognition in aqueous media. New binding studies provide further insights into the cation-π interaction and related phenomena. *J. Am. Chem. Soc.*, 115:9907–9919, 1993.

[12] J.L. Sussman, M. Harel, F. Frolow, C. Oefner, A. Goldman, L. Toker, and I. Silman. Atomic structure of acetylcholinesterase from *Torpedo californica*: A prototypic acetylcholine-binding protein. *Science*, 253:872–879, 1991.

[13] R.A. Kumpf and D.A. Dougherty. A mechanism for ion selectivity in potassium channels: Computational studies of cation-π interactions. *Science*, 261:1708–1710, 1993.

[14] J. W. Caldwell and P. A. Kollman. Cation-π interactions: Nonadditive effects are critical in their accurate representation. *J. Am. Chem. Soc.*, 117:4177–4178, 1995.

[15] S. Mecozzi, A. P. West Jr., and D. A. Dougherty. Cation-π interactions in simple aromatics. Electrostatics provide a predictive tool. *J. Am. Chem. Soc.*, 118:2307–2308, 1996.

[16] S. Mecozzi, A. P. West Jr., and D. A. Dougherty. Cation-π interactions in aromatics of biological and medicinal interest: Electrostatic potential surfaces as a useful qualitative guide. *Proc. Natl. Acad. Sci. USA*, 93:10566–10571, 1996.

[17] E. Cubero, F. J. Luque, and M. Orozco. Is polarization important in cation-π interactions? *Proc. Natl. Acad. Sci. USA*, 95:5976–5980, 1998.

Chapter 2 From Ab Initio Quantum Mechanics to Molecular Neurobiology: A Cation-π Interaction in the Nicotinic Acetylcholine Receptor

2.1 Introduction

The nicotinic acetylcholine receptor (nAChR) is the prototype ligand-gated ion channel [1–3]. It is the longest known and most well studied member of a superfamily that includes receptors for serotonin, γ-aminobutyric acid, and glycine. There are two main classes of nicotinic acetylcholine receptors, neuronal and muscle. Neuronal nicotinic acetylcholine receptors represent a large family of receptors that function in the central nervous system and play an important role in nicotine addiction. In contrast, the muscle-type nAChR, which is the primary focus of this chapter, functions at the neuromuscular junction. When an action potential reaches the end of a motor neuron, a pulse of acetylcholine (ACh) is released into the synaptic cleft. ACh diffuses across the synaptic cleft where it may bind to the nicotinic acetylcholine receptor. When two molecules of ACh bind to the receptor, a non-selective cation channel within the receptor opens, which allows Na^+ and K^+ to flow into the post-synaptic cell and causes the muscle to contract. Throughout this process, acetylcholinesterase, an enzyme located on the post-synaptic cell, rapidly hydrolyzes acetylcholine to produce choline and acetate, thereby terminating the chemical signal.

The nicotinic acetylcholine receptor is a remarkable molecular machine. Upon binding two small molecules, the receptor converts a chemical signal into an electrical one. As chemists, we would like to understand both the structure and function of the nicotinic acetylcholine receptor at the chemical scale. In doing so, we would like to answer several fundamental questions: What does the receptor look like? How does ACh bind? How is this binding converted into a conformational change? The major focus of this chapter is to answer the question of how acetylcholine binds to the nicotinic acetylcholine receptor. The studies in this chapter were carried out in collaboration with Wenge Zhong, who performed all of the experimental work described.

2.2 Characterizing the nAChR at the Molecular Level

Because the nicotinic acetylcholine receptor is an integral membrane protein, it has resisted structural characterization at the molecular level. Standard high-resolution structural methods, such as x-ray crystallography and nuclear magnetic resonance spectroscopy, are challenged to provide molecular scale information about integral membrane proteins. This is often due in part to the lack of sufficient quantities of material to perform structural studies. In the case of the nAChR, Nature has provided a plentiful source of protein that can be used for structural and biochemical studies. The electric organ of certain species of rays and eels, such as *Torpedo californica*, contains a high density of nAChRs. As a result, much is known about this receptor from both classical biochemistry as well as groundbreaking electron microscopy studies performed in the laboratory of Nigel Unwin [4–7].

Our present view of the receptor suggests that it is a large protein (MW 290 kD) comprising five homologous subunits. The subunits are thought to be arranged in a pentagonal array around a central pore that forms the ion channel. The embryonic form of the receptor has a subunit stoichiometry of $\alpha_2, \beta, \gamma, \delta$. The arrangement of these subunits around the pore is still the subject of debate, but the most commonly advocated arrangement is a clockwise arrangement of $\alpha, \gamma, \alpha, \delta, \beta$ when viewed from the extracellular side of the membrane as shown in Figure 2.1.

2.2.1 Structural Data

The three-dimensional structure of the nicotinic acetylcholine receptor has been determined to 9 Å resolution [5] (and 4.6 Å in projection) [7] using electron microscopy. Although these experiments do not reveal the atomic positions of either the protein backbone or individual side chains, they do offer insight into the global layout of the receptor with respect to the lipid membrane. Several structural features are highlighted in Figure 2.1. The receptor is approximately 120 Å long and 85 Å across at

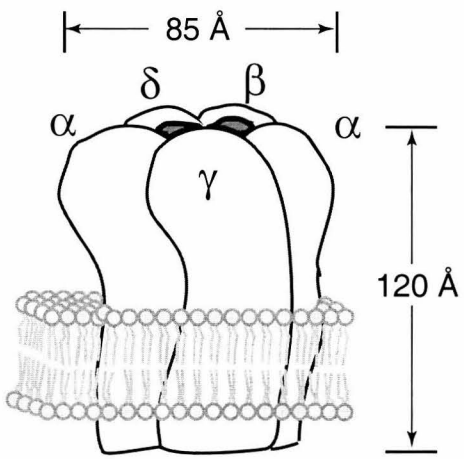


Figure 2.1: Structural features of the nAChR taken from the work of Unwin [5], with the subunit arrangement taken from Hucho *et al.* [3]

its widest point. The extracellular domain extends approximately 76 Å above the lipid bilayer, and the agonist binding sites, which are believed to be located primarily on the α-subunits, are proposed to lie approximately 30 Å above the lipid bilayer [5]. It has been proposed that the gate of the receptor is a hydrophobic plug consisting of a ring of highly conserved leucine residues located within the membrane [8]. These residues, which are proposed to be located in the second membrane-spanning region (known as M2) of the individual subunits, are believed to be located over 50 Å away from the putative agonist binding sites (Figure 2.2).

2.2.2 Sequence Landmarks

Although protein isolated from *Torpedo* has provided many details about the structure and function of the nicotinic acetylcholine receptor, it serves as only a starting point. As physical organic chemists, we would like to perform structure-function studies on the receptor. These studies were enabled by the cloning of the subunits of the receptor by Numa [9–12] and Claudio [13], and the subsequent heterologous expression of the receptor in *Xenopus* oocytes by Numa [14]. The sequences reveal that the subunits of the receptor are highly homologous. Hydropathy analysis predicts that each subunit contains a large (*ca.* 200 amino acid) extracellular domain, followed by four membrane-spanning regions (designated M1-M4, Figure 2.3).

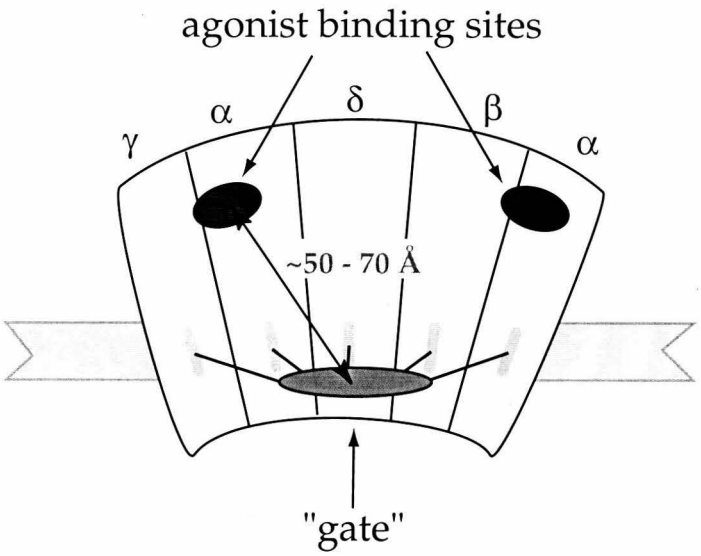


Figure 2.2: Cutaway view of the nAChR. The agonist binding sites are postulated to be located over 50 Å away from the gate of the ion channel.

In the extracellular domain of each subunit, there are two cysteine residues that form a 15-residue disulfide-loop. This signature disulfide cys-loop is a hallmark of the ligand-gated ion channel superfamily and is present in all nicotinic receptors, as well as the receptors for glycine, serotonin, and γ-aminobutyric acid (GABA) [2]. In addition to the cys-loop, there is a vicinal disulfide linkage that is unique to α-subunits of the nAChR that is known to be near the ACh binding site. Although binding of acetylcholine is localized to the extracellular domain, the gate, which controls the passage of ions, is believed to be located in the membrane. Extensive studies of the receptor suggest that the M2 region of each subunit lines the ion channel, and that one residue in particular, leucine 9′, forms the gate of the channel [15, 16], although this model is not universally accepted [17, 18].

2.2.3 The Agonist Binding Site

The agonist binding sites of the nicotinic acetylcholine receptor are believed to be located at the interfaces of the α and γ subunits and the α and δ subunits [19–22]. Binding of acetylcholine to these sites is cooperative and leads to channel opening [23]. A fundamental goal of this work is to identify which residues bind ACh,

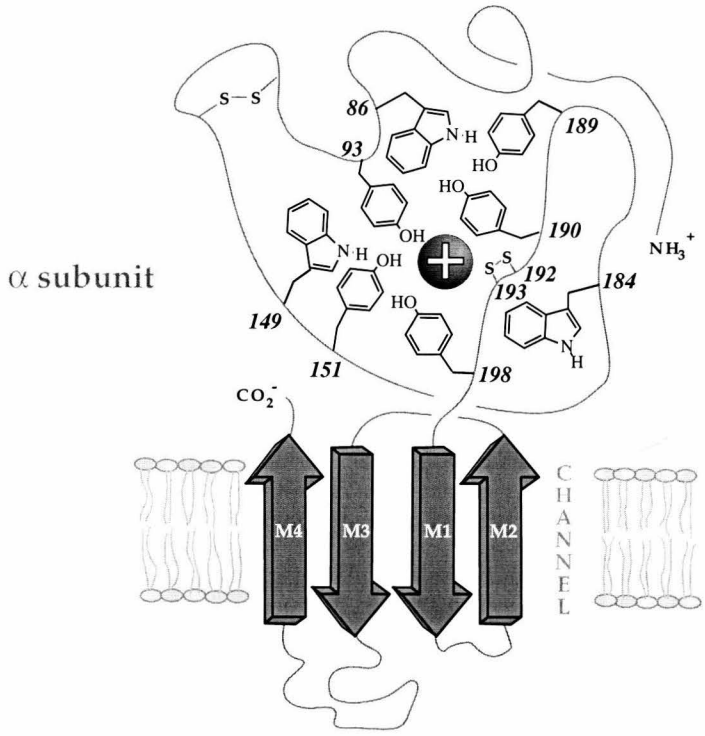


Figure 2.3: Schematic view of the α subunit of the nAChR. The other subunits are highly homologous and are believed to share the same overall topology. The aromatic residues highlighted have been affinity-labeled (see text).

and through what forces. Although no direct high-resolution structural data exist for the nAChR, pioneering labeling studies carried out on the *Torpedo* nAChR have implicated several amino acids that may be important in the binding of ACh. Using a variety of labeling reagents, such as dimethylaminobenzene diazonium tetrafluoroborate (DDF) [24, 25], 4-(N-maleimido)-benzyltrimethylammonium (MBTA) [26], nicotine [22], *d*-tubocurarine [27], lophotoxin [28], and acetylcholine mustard [29], several residues that are likely involved in the binding of ACh have been identified, including αTrp86, αTyr93, αTrp149, αTyr151, αTyr190, αCys192, αCys193, αTyr198, γTrp55, and δTrp57. Not surprisingly, all labeled residues within the α-subunits are highly conserved and some (αTrp86, αTyr151, and αTyr198) have been called structural canonical residues. What is at first glance surprising is that the majority of labeled residues are aromatic. Conventional wisdom might suggest that to bind a cationic agonist such as acetylcholine, there must be an anionic residue. However, no anionic residues were directly labeled. This large number of aromatic amino acids led

Dougherty and Stauffer to propose that cation-π interactions might be important in the binding of acetylcholine to the nicotinic acetylcholine receptor [30]. The remainder of this chapter describes our attempts to test this proposal.

2.3 Cation-π Interactions and Structure-Function Studies

One of the most time-tested methods of physical organic chemistry is the determination of a structure-function relationship. Put simply, how do changes in molecular structure effect the function of a system? To answer this question, we must be able to do three things. We must first be able to perform structural changes on the nicotinic receptor. Second, we must be able to measure the effect of these structural changes on the function of the receptor. The third requirement is that we must be able to correlate the results from the structure and function studies to some known molecular property, in this case, cation-π binding ability.

2.3.1 Site-Directed Mutagenesis in Molecular Neurobiology—Varying Structure

Arguably the most powerful weapon in the arsenal of the molecular neurobiologist is site-directed mutagenesis – changing one amino acid to another. By altering the molecular structure of an amino acid side chain, one can attempt to correlate the structural change with a functional change. Although site-directed mutagenesis is quite versatile, it is limited by the fact that each amino acid may only be changed to one of the other nineteen translationally-incorporated amino acids. As physical organic chemists, we would like finer control over the amino acid substitutions. Fortunately, work from the laboratories of Schultz [31, 32], Chamberlain [33–35], and Hecht [36] has provided us with a method to expand the genetic code and site-specifically incorporate unnatural amino acids into proteins expressed in vitro using a technique known as nonsense codon suppression. In the nonsense codon suppres-

sion method, a mutant mRNA is produced that contains a stop codon at the site of interest. Normally, this codon would signal the termination of translation because there is no tRNA molecule with an anticodon that recognizes the stop codon. To circumvent this, a tRNA with an anticodon that recognizes the stop codon is produced in vitro. This "suppressor" tRNA is chemically aminoacylated in vitro with the desired amino acid. By combining the mutant mRNA and aminoacyl-tRNA in an appropriate translation system, proteins containing the desired amino acid at the site of interest can be produced. Nonsense codon suppression is a powerful technique and has been used to site-specifically incorporate both natural and unnatural amino acids into proteins expressed in vitro [31–36].

Unfortunately, as originally developed, the nonsense codon suppression method is not generally applicable to problems in molecular neurobiology. This is largely due to the fact that the proteins of molecular neurobiology are often complex, heavily-processed, multi-subunit integral membrane proteins. Although these proteins may be expressed in vitro, they may not fold or assemble correctly, and thus they cannot be assayed in a proper biological context. Ideally, one would like to assay the function of these proteins in a biologically relevant context – within a living cell. In the past several years, the Dougherty and Lester laboratories have adapted the nonsense codon suppression method for use in a living cell, specifically a *Xenopus* oocyte [37–39]. Today, it is possible to site-specifically incorporate unnatural amino acids into proteins expressed in vivo. We now have unprecedented structural control over the molecular properties of individual amino acid side chains. Over 60 natural and unnatural amino acids have been incorporated into ion channel proteins expressed in a living cell.

2.3.2 Electrophysiological Recording – Measuring Function

Site-directed mutagenesis is a powerful method of making structural changes in a protein. However, we must also be able to assay the functional effects of such changes. Because the nicotinic acetylcholine receptor is an ion channel, one way to assay its function is to measure the flow of ions through the channel. This can be done using the

two-electrode voltage clamp technique. In the two-electrode voltage clamp technique, a cell expressing the protein of interest is pierced with two electrodes. The first electrode (known as the voltage electrode) serves to monitor the membrane potential of the cell. The membrane potential can be set to an arbitrary value and 'clamped' (i.e., maintained at a constant value). The second electrode (known as the current electrode) delivers current to the cell such that the membrane potential remains at the clamped value. If the membrane potential deviates from the clamped value due to the opening or closing of ion channels, the current electrode must deliver current to compensate for this change to maintain the clamped voltage. The current delivered is equal in magnitude, but opposite in sign, to the current passed by the ion channels. Thus, by measuring the total amount of current required to maintain a voltage clamp (the whole-cell current), one can determine the total amount of current passing through the open channels. The nicotinic acetylcholine receptor is a ligand-gated ion channel. In the absence of an agonist, such as acetylcholine, the probability of the channel opening is quite low. Application of ACh increases the probability that a conformational change will occur that allows ions to flow across the membrane. The whole-cell current of *Xenopus* oocytes expressing the nAChR can thus be measured in response to varying concentrations of ACh. The concentration of agonist that elicits half of the maximal response is known as the EC_{50} for that agonist. It is important to note that although EC_{50} measures the sensitivity of a receptor for a given agonist, it is not a binding constant. Rather, EC_{50} is a composite value that incorporates the effects of both agonist binding and channel gating. Nonetheless, small variations in EC_{50} caused by mutations often correlate well with the free energy change of binding of the ligand [40]. Thus, by measuring the EC_{50} of ACh for a series of mutant receptors, we can quantitatively evaluate receptor function.

2.3.3 Correlating Structure and Function

It is now clear that we can vary the structure of the nAChR by using nonsense codon suppression. In addition, we can assay the effects of the mutations by measuring the

sensitivity of the mutant receptors to ACh by recording the EC_{50}. The remaining challenge is to correlate changes in the structure and function of the receptor to a change in the cation-π binding ability of an individual amino acid side chain. Fortunately, Anthony West and Sandro Mecozzi convincingly demonstrated that variations in the cation-π binding abilities of a structurally-related series of aromatics could be accounted for by variations in the electrostatic potential surfaces of the aromatics [41, 42]. Thus, we can modify the electrostatic potential (and thus the cation-π binding ability) of an amino acid via organic synthesis; we can evaluate the variations using *ab initio* quantum mechanical calculations; and we can measure the changes in EC_{50} for ACh for receptors containing the mutant side chains. If the changes in EC_{50} for a series of closely-related aromatics at a particular site in the receptor follow a consistent electrostatic trend, we will have evidence that the particular amino acid is involved in a cation-π interaction with acetylcholine.

2.4 Previous Studies of Possible Cation-π Interactions in the nAChR

As noted in Section 2.2, several aromatic amino acids have been implicated in the binding of acetylcholine by the nAChR, including αTrp86, αTyr93, αTrp149, αTyr151, αTrp184, αTyr190, αTyr198, and the homologous γTrp55 and δTrp57. Previous work from the Dougherty group explored the possibility that the tyrosines at positions α93, α190, and α198 might bind ACh through a cation-π interaction [37]. Using the in vivo nonsense suppression method, it was determined that the hydroxyl group of αTyr93 functions as a hydrogen bond donor, and that αTyr198 serves as a steric place holder in the receptor. αTyr190 was found to be very sensitive to subtle mutations, but no consistent structure-function relationship could be derived [37]. Perhaps surprisingly, none of the three tyrosines studied appeared to bind to ACh through a cation-π interaction. In addition to the tyrosine residues, there are several conserved tryptophans that have been affinity labeled. It might be anticipated that a tryptophan would con-

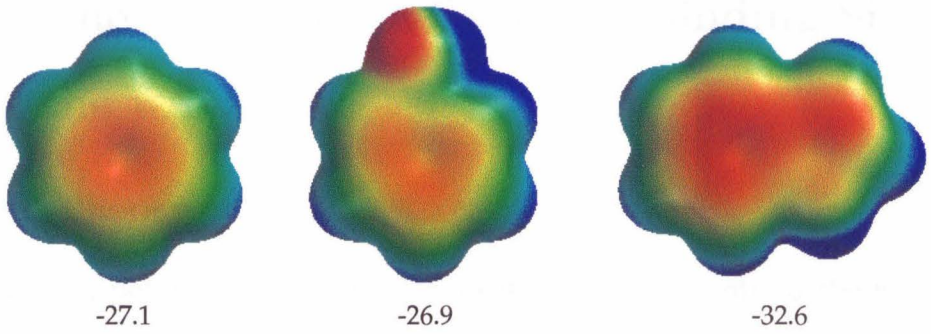


Figure 2.4: Calculated HF/6-31G** electrostatic potential surfaces and Na^+ binding energies (see text) of the aromatic side chains of phenylalanine (benzene), tyrosine (phenol), and tryptophan (indole). Red represents a region of negative electrostatic potential, blue positive. *(Scale: -25 to +25 kcal/mol)* .

stitute a cation-π binding site because the indole side chain of tryptophan provides the most potent cation binding site among the aromatic amino acids (Figure 2.4) [42].

There are relatively few studies of the roles of the conserved tryptophans in the nAChR. Both White [43] and Changeux [44] have demonstrated that γTrp55 and δTrp57 are important binding determinants of both ACh and *d*-tubocurarine in the receptor from *Torpedo*. In addition, Changeux has shown that a Trp→Phe mutation at position α148 in the α7 neuronal receptor (analogous to α149 in the muscle receptor) increases the EC_{50} for ACh nearly one hundred-fold [44]. Finally, in an NMR study of peptide fragments from the α-subunit, Fraenkel reported that the quaternary ammonium group of ACh contacts the indole side chain of α184 [45]. Although these studies suggest the possibility of cation-π interactions between ACh and the tryptophans in the nAChR, they do not constitute proof that such an interaction exists. The remainder of this chapter describes our efforts to find a cation-π binding site within the nicotinic acetylcholine receptor.

2.5 The Search for a Cation-π Binding Site within the nAChR

2.5.1 Theory

With several potential candidates for a cation-π binding site in the nAChR, we set out to predict the cation-π binding ability of a series of tryptophan derivatives that could be used as replacements in a nonsense codon suppression experiment. Previous studies by Anthony West and Sandro Mecozzi have shown that the cation-π binding ability of a simple aromatic can be quantitatively determined by using high-level *ab initio* calculations [41, 42]. To predict the cation-π binding ability of an aromatic, the binding energy between the aromatic ring and a probe cation, such as Na^+, is determined using an *ab initio* calculation. Earlier studies have suggested that Hartree-Fock calculations with a medium-sized basis set, such as 6-31G**, can accurately reproduce experimentally-determined binding energies of aromatics and non-polarizable cations [46]. In addition, incorporation of the effects of electron correlation using post-Hartree-Fock methods, such as Møller-Plesset (MP) perturbation theory, does not significantly improve the agreement between experiment and theory in systems containing non-polarizable cations. Thus, we used Hartree-Fock theory with the 6-31G** basis set to calculate the Na^+ binding energy for the series of indole analogs shown in Figure 2.5.

To perform the calculations, a Na^+ ion was placed over the ring of interest, and the geometry of the complex was fully optimized using Gaussian 94 or Spartan. The binding energies were calculated using the supermolecule method and are not corrected for basis set superposition error (BSSE) [47], which is known to be small in these simple systems. The calculated binding energies of these systems are shown in Table 2.1.

There are several items worthy of explanation in Table 2.1. Perhaps most surprising is the extraordinary cation-π binding ability of the indole side chain of tryptophan. Of the 22 indole analogs shown in Table 2.1, only three bind cations more strongly

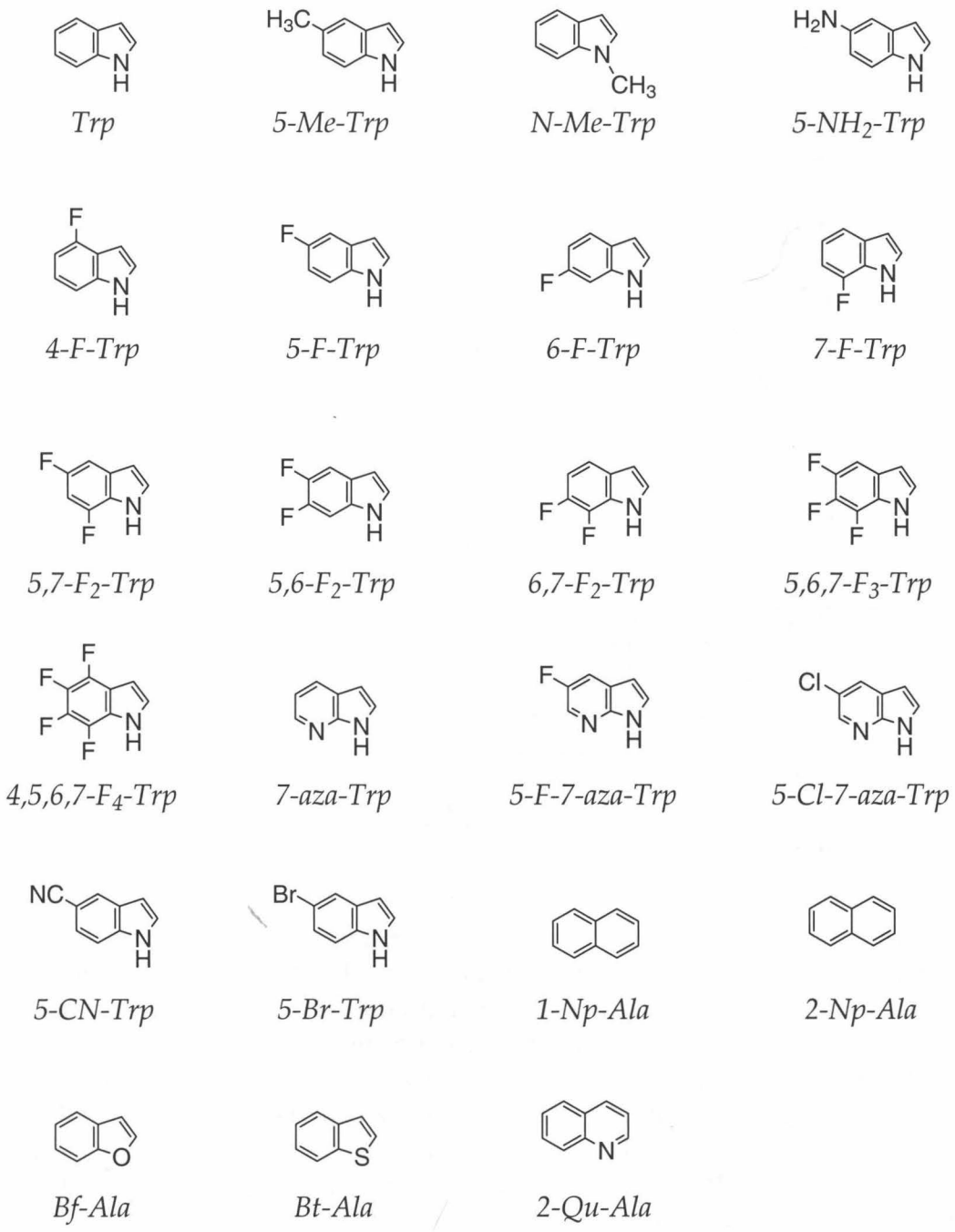


Figure 2.5: Side chain structures of a series of tryptophan analogs. The derivatives are named as amino acids; thus indole is considered Trp.

Aromatic	Cation-π binding of 6-ring (kcal/mol)	Cation-π binding of 5-ring (kcal/mol)
Trp	32.6	29.0
5-Me-Trp	33.4	n.d.
N-Me-Trp	33.7	n.d.
5-NH_2-Trp	36.4	n.d.
4-F-Trp	27.9	n.d.
5-F-Trp	27.5	25.4
6-F-Trp	27.4	n.d.
7-F-Trp	28.2	n.d.
5,7-F_2-Trp	23.3	21.9
5,6-F_2-Trp	23.2	n.d.
6,7-F_2-Trp	23.4	n.d.
5,6,7-F_3-Trp	18.9	18.9
4,5,6,7-F_4-Trp	14.4	15.6
7-aza-Trp†	26.0	25.0
5-F-7-aza-Trp†	21.2	n.d.
5-Cl-7-aza-Trp†	20.8	n.d.
5-CN-Trp	21.5	19.1
5-Br-Trp ‡	27.8	n.d.
1-Np-Ala, 2-Np-Ala$^{\triangle}$	28.9	n.a.
Bf-Ala	26.9	22.3
Bt-Ala	26.9	n.d.
2-Qu-Ala	25.8	n.a.

Table 2.1: Cation-π binding energies for the aromatic derivatives shown in Figure 2.5. The cation-π binding is defined as the negative of the binding energy (kcal/mol) of a Na^+ ion to the aromatic ring determined from a HF/6-31G** calculation. In cases where the 5-membered ring of the aromatic might compete with the 6-membered ring for cation binding, the binding energy to the 5-membered ring is presented. † The true minima for the aza-Trp derivatives have the cation bound to the pyridine nitrogen. The values reported in the table in some cases had to be constrained to the ring axis. ‡ The value for 5-Br-Trp is determined from an HF/6-311G** calculation (see text). $^{\triangle}$ 1-Np-Ala and 2-Np-Ala are listed together because they have the same side chain. The Np-Ala values were calculated by Sandro Mecozzi [41]. n.a. – not applicable, n.d. – not determined.

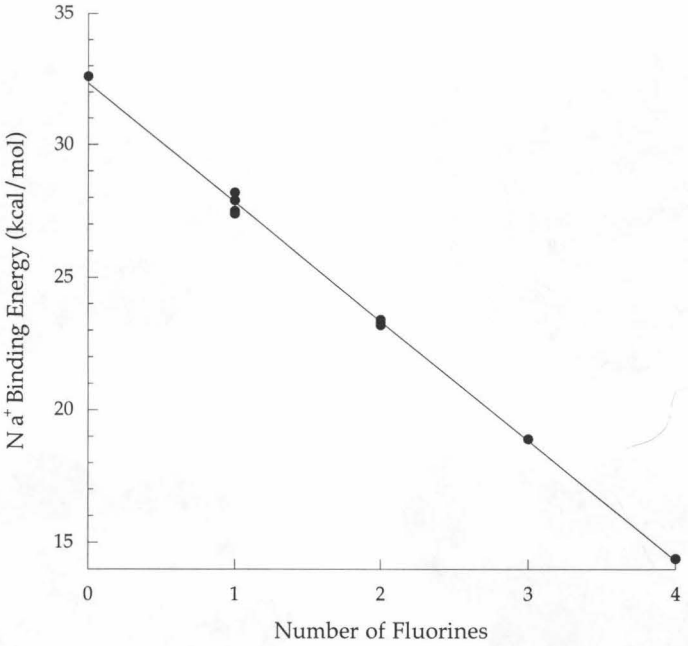


Figure 2.6: The calculated HF/6-31G** Na^+ binding energies of a series of indole derivatives fluorinated on the 6-membered ring. Each fluorine reduces the Na^+ binding energy by 4.5 kcal/mol. ($R = 0.99$).

than indole itself. The most potent of these, 5-NH_2-indole, might be expected to easily oxidize and be less stable. Thus, it would appear that Nature has chosen a particularly strong, yet chemically stable, cation-binder in the indole side chain of tryptophan. Although it is difficult to strengthen the cation-π binding ability of the indole ring, it is quite easy to weaken it. Consistent with previous studies, electron-withdrawing substituents, such as fluorine and cyano groups, substantially weaken the cation-π binding ability of an aromatic ring [41]. In addition, increased fluorination directly correlates with decreased cation-π binding ability. As shown in Figure 2.6, each additional fluorine reduces the cation-π binding ability of the ring by a constant 4.5 kcal/mol.

The data in Table 2.1 clearly suggest a broad range of cation-π binding abilities for the various indole analogs. Thus, it should be possible to synthesize a series of tryptophan derivatives that have widely differing cation-π binding abilities. However, to establish whether a cation-π interaction is occurring, we need to show a correlation between the predicted cation-π binding ability of a given amino acid and the function

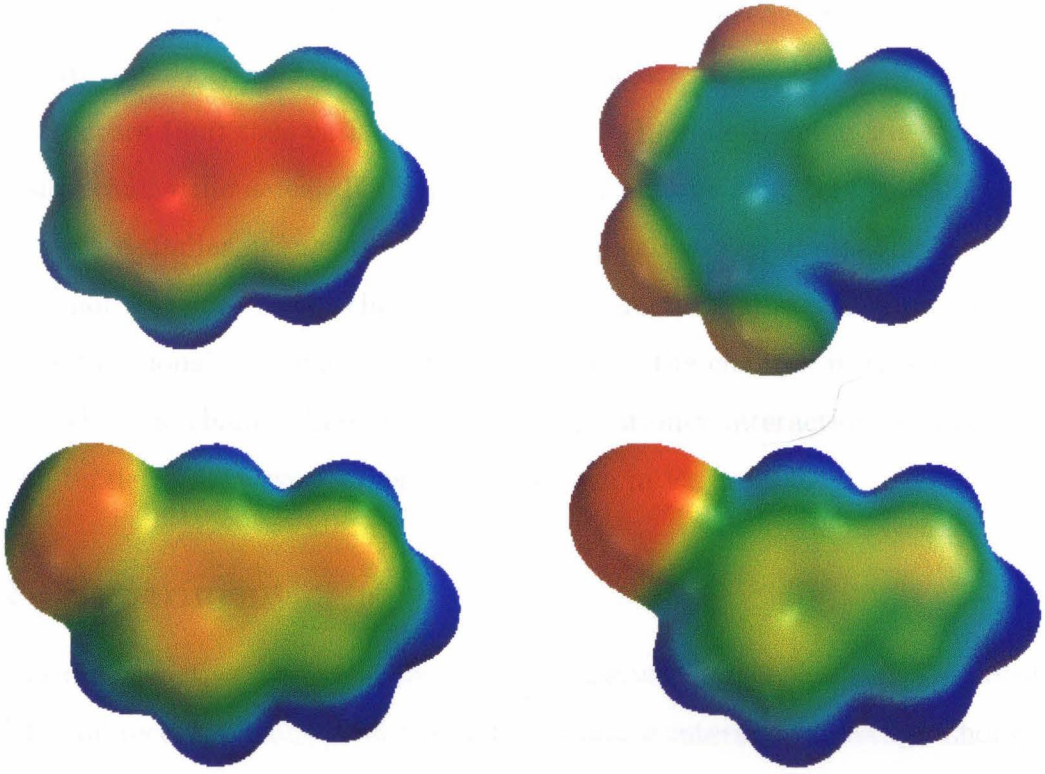


Figure 2.7: Calculated electrostatic potential surfaces of the side chains of Trp, 4,5,6,7-F_4-Trp, 5-Br-Trp, and 5-CN-Trp. It is clearly seen that the introduction of electron-withdrawing groups, such as fluorine and cyano groups, lowers the electrostatic potential above the rings. In addition, this figure highlights the similar size of the bromine and cyano substituents. The surfaces are calculated at the HF/6-31G** level for Trp, 4,5,6,7-F_4-Trp, and 5-CN-Trp, and at the HF/6-311G** level for 5-Br-Trp. Justification for the use of the 6-311G** basis set appears in an appendix to this chapter. *(Scale: -25 to +25 kcal/mol).*

of a receptor containing that amino acid. Four entries in Table 2-1 are particularly useful in this regard: Trp, 4,5,6,7-F_4-Trp, 5-Br-Trp, and 5-CN-Trp. The calculated electrostatic potential surfaces of these side chains are shown in Figure 2.7.

Figure 2.7 highlights several important comparisons between the various side chains. It is clearly seen, both qualitatively and quantitatively (Table 2.1), that the addition of four fluorines reduces the cation-π binding ability of the indole side chain of tryptophan. Fluorine is generally regarded as a conservative steric substitute for a hydrogen atom. Thus, the substitution of four fluorines alters the electronic properties on the ring dramatically, while leaving the steric properties unperturbed. In a similar manner, 5-Br-Trp and 5-CN-Trp are nearly isosteric, yet their cation-π

binding abilities are quite different. Thus, by mutating the conserved tryptophans in the nAChR to only these 3 Trp derivatives, we can screen the effects of varying electronic structure (and thus cation-π binding ability), while keeping the steric profile of the side chains the same. If the properties of mutant receptors containing these derivatives are similar, it suggests that changes in cation-π binding ability of the side chains do not correlate with a change in receptor function. However, if mutant receptors show functional differences that correlate with the changes in cation-π binding ability of the side chains, there is evidence of a cation-π interaction, and additional tests can be performed using other analogs shown in Figure 2.5.

2.5.2 Experiment

To determine whether one or more of the conserved tryptophans in the nicotinic acetylcholine receptor binds ACh through a cation-π interaction, Wenge Zhong synthesized several of the Trp analogs outlined in Figure 2.5. Using the nonsense codon suppression method, he incorporated Trp and the three analogs, 4,5,6,7-F_4-Trp, 5-Br-Trp, and 5-CN-Trp at several sites within the nAChR. The EC_{50} for ACh for the wild-type Trp and each of the three Trp derivatives was measured at several sites, α86, α149, α184, γ55, and δ57, and the results are reported in Table 2.2.

	EC_{50} (μM)			
Side chain	α86	α149	α184	γ55/δ57
Trp	50	50	50	50
5-Br-Trp	53	88	48	23
5-CN-Trp	61	4750	46	25
4,5,6,7-F_4-Trp	45	2700	62	92

Table 2.2: EC_{50} values for ACh for several Trp mutants at α86, α149, α184, γ55, and δ57. There are only substantial variations at α149, suggesting a possible cation-π interaction between ACh and αTrp149. The mutations in the δ and γ subunits are expressed simultaneously, so each agonist binding site contains the mutant Trp. Because the EC_{50} is so high for some mutations, it is difficult to measure accurately. To make these receptors more responsive to ACh, an additional mutation, Leu 9′→Ser was introduced and the values were extrapolated. For full details of this process, see [48].

The results in Table 2.2 suggest that at three of the four Trp sites (α86, α184,

Aromatic	Cation-π binding (kcal/mol)	EC_{50} (μM)
Trp	32.6	50
5-Me-Trp	33.4	49
N-Me-Trp	33.7	95
5-NH_2-Trp	36.4	280
4-F-Trp	27.9	56
5-F-Trp	27.5	200
6-F-Trp	27.4	48
5,7-F_2-Trp	23.3	550
5,6,7-F_3-Trp	18.9	1400
4,5,6,7-F_4-Trp	14.4	2700
7-aza-Trp	26.0	130
5-CN-Trp	21.5	4750
5-Br-Trp	27.8	88
1-Np-Ala	28.9	51
2-Np-Ala	28.9	82
Bt-Ala	26.9	174

Table 2.3: Calculated cation-π binding abilities and experimentally-measured EC_{50} values for ACh for various tryptophan derivatives incorporated into the nAChR.

α55/α57), large variations in the cation-π binding ability of the amino acid side chain do not significantly alter the sensitivity of the receptor for ACh. Thus, there is little evidence to suggest a cation-π interaction at these sites. However, at α149, there is a large change in the EC_{50} for ACh that does correlate with the cation-π binding ability of the amino acid side chains. In particular, the EC_{50} values for both 5-CN-Trp and 4,5,6,7-F_4-Trp are substantially higher than the values for either 5-Br-Trp or Trp. This is consistent with weaker ACh binding leading to a higher EC_{50}. To further investigate the possibility of a cation-π interaction between the quaternary ammonium group of ACh and the side chain of αTrp149, Wenge Zhong synthesized several other Trp analogs, incorporated them into the nAChR, and measured the EC_{50} for ACh. The results of these experiments are shown in Table 2.3.

Although the results for 5-CN-Trp are highly suggestive of a cation-π interaction between αTrp149 and ACh, an even more compelling correlation is observed for the series of fluorinated Trp derivatives shown in Table 2.3. In Figure 2.8, the experimentally-determined $\log(EC_{50})$ values for a series of fluorinated Trp derivatives

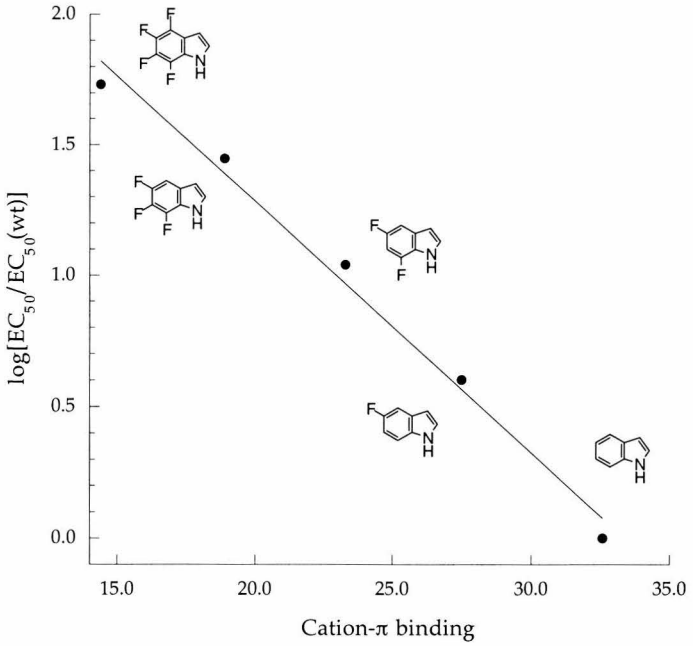


Figure 2.8: Plot of log[EC_{50}/EC_{50}(wild type)] vs. quantitative cation-π binding ability of Trp and its fluorinated derivatives expressed at α149. Data are taken from Table 2.3 and fit the line $y = 3.2 - 0.096x, R = 0.99$. The log[$EC_{50}$] values are used to place the $y-$axis on an energy scale.

is plotted against the calculated cation-π binding abilities of the derivatives.

Figure 2.8 incorporates two very different measurements. On one axis are the predicted cation-π binding abilities for the various amino acid side chains calculated using *ab initio* quantum mechanics. On the other axis are experimental measurements performed on a molecular weight 290,000 protein expressed in a living cell. Nonetheless, a striking correlation is observed between these two very different measurements. It should again be stressed that the EC_{50} for ACh is not a binding constant. Rather, it incorporates the effects of both binding and gating. In Figure 2.8, we do not assume that gating is unimportant, but rather that the changes in EC_{50} are primarily due to changes in the binding affinity for ACh. To the extent that EC_{50} is proportional to a binding constant, log[EC_{50}] is the appropriate quantity to correlate with a binding energy. The EC_{50} change observed on going from Trp to 4,5,6,7-F_4-Trp represents an energy difference of 2.4 kcal/mol of binding energy. Since it is unlikely that this mutation substantially alters the structure of the receptor, it is reasonable to assume that this energy change is directly related to the binding of ACh. The energy dif-

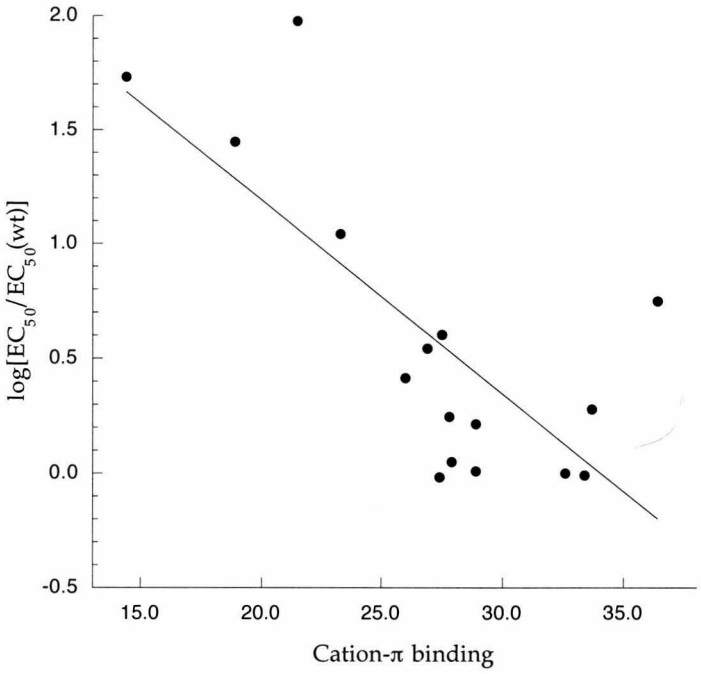


Figure 2.9: Plot of log[EC_{50}/EC_{50}(wild type)] vs. quantitative cation-π binding ability of all Trp derivatives expressed at α149. Data are taken from Table 2.3. The outliers correspond to 5-CN-Trp and 5-NH_2-Trp. The correlation coefficient for the line shown is 0.74. Removing the outliers improves the correlation to $R = 0.90$.

ference of 2.4 kcal/mol is quite large for a noncovalent interaction and thus suggests that ACh and the Trp side chain are interacting directly. In addition to the data from the fluorinated tryptophans, a plot containing all of the data in Table 2.3 reveals a similar trend as shown in Figure 2.9.

The outlying data points in Figure 2.9 correspond to 5-CN-Trp and 5-NH_2-Trp. It is not clear why these particular amino acids are outliers, but there is a possibility that 5-NH_2-Trp is protonated or oxidized and thus leads to a higher than expected EC_{50}. Although there is scatter in the data, the trend is clear that a decrease in cation-π binding ability of an amino acid side chain correlates with an increase in EC_{50} for ACh.

Until this point, we have assumed that the ACh is in contact with the 6-membered ring of the Trp. This is a reasonable assumption given that it is a more potent cation-π binder than the 5-membered ring. However, there is a possibility that ACh is in fact binding to the 5-membered ring of the Trp. A plot of the EC_{50} data taken from

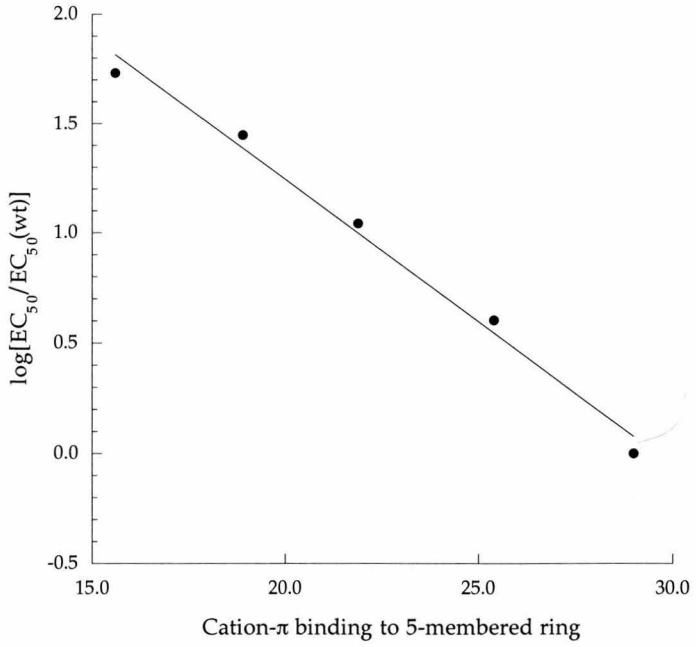


Figure 2.10: Plot of $\log[EC_{50}/EC_{50}(\text{wild type})]$ vs. quantitative cation-π binding ability of the 5-membered ring of Trp and its fluorinated derivatives expressed at α149. Data are taken from Table 2.3 and fit the line $y = 3.84 - 0.13x, R = 0.99$.

Table 2.3 vs. the calculated cation-π binding abilities of the 5-membered rings of Trp and its fluorinated derivatives (taken from Table 2.1) is shown in Figure 2.10.

Indeed an excellent correlation is also observed when considering the 5-membered rings of Trp and its fluorinated derivatives. However, if the ACh were in contact with the 5-membered ring of the Trp, it might be expected that phenylalanine, with a side chain aromatic that aligns with the 5-membered ring of the Trp, would make a good substitute for Trp. Experimentally, this is not the case. Changeux has observed that an analogous Trp→Phe mutation in the α7 neuronal nAChR markedly increased the EC_{50} for ACh [44], and we find that an αTrp149→Phe mutation increases the EC_{50} for ACh approximately 100-fold. Thus it is likely that the quaternary ammonium group of ACh is in contact with the 6-membered ring of the Trp. However, one cannot exclude the possibility that the 6-ring of Trp positions the side chain such that ACh can bind to the 5-ring, whereas Phe, lacking a second place-holder ring, might not be able to bind as well. An amino acid such as 4,5,6,7-tetrahydro-Trp might be able to experimentally distinguish the two possibilities.

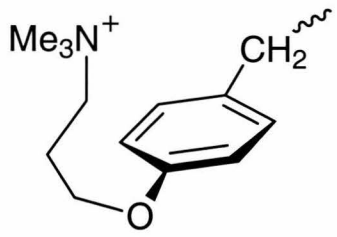


Figure 2.11: Side chain structure of the tethered agonist Tyr-O-3Q.

Finally, inspired by the information derived from this study, Wenge Zhong incorporated into position α149 an amino acid with a quaternary ammonium group tethered to its side chain (Tyr-O-3Q, Figure 2.11).

Incorporation of Tyr-O-3Q at position α149 produces a constitutively active receptor. In the absence of ACh, standing inward currents that can be inhibited by nAChR open-channel blockers are seen in *Xenopus* oocytes expressing the mutant channel. The side chain of Tyr-O-3Q positions the quaternary ammonium group such that it can mimic the binding of ACh to the 6-membered ring of αTrp149. This result and others further suggest that ACh binds to αTrp149 of the nAChR via a cation-π interaction.

2.6 Summary

The work presented in this chapter illustrates the power of combining theoretical and physical organic chemistry with modern neurobiological techniques to provide chemical-scale information of a complex molecular machine. By combining *ab initio* quantum mechanical calculations, organic synthesis, and molecular neurobiology, we provide compelling evidence that the quaternary ammonium group of acetylcholine binds to the nAChR through a cation-π interaction with αTrp149. In doing so, we provide some of the highest-resolution structural information known about this receptor. In addition, we demonstrate the power of the nonsense codon suppression method to apply the principles of physical organic chemistry to complex biological systems.

Aromatic	6-31G** minimum energy of monomer (Hartrees)	6-31G** minimum energy of Na^+ complex with 6-membered ring (Hartrees)
Trp	–361.4829178	–523.1941820
5-Me-Trp	–400.5221750	–562.2346708
N-Me-Trp	–400.5146125	–562.2275997
5-NH_2-Trp	–416.513102	–578.2304357
4-F-Trp	–460.3330449	–622.0367912
5-F-Trp	–460.3316404	–622.0347153
6-F-Trp	–460.3327604	–622.0357012
7-F-Trp	–460.3308288	–622.0351006
5,7-F_2-Trp	–559.1788382	–720.8752736
5,6-F_2-Trp	–559.1736047	–720.8691313
6,7-F_2-Trp	–559.1723013	–720.8688588
5,6,7-F_3-Trp	–658.0126328	–819.7020758
4,5,6,7-F_4-Trp	–756.8518351	–918.5340764
7-aza-Trp	–377.4819407	–539.1826815
5-F-7-aza-Trp	–476.3267973	–638.0199006
5-Cl-7-aza-Trp	–836.3784228	–998.0708002
5-CN-Trp	–453.2199701	–614.9135390
5-Br-Trp †	–2933.331860	–3095.036333
Bf-Ala	–381.2991455	–543.0012739
Bt-Ala	–703.9623685	–865.6646036
2-Qu-Ala	–399.3623064	–561.0627336

Table 2.4: Total energies of monomers and 6-ring Na^+ complexes. † 5-Br-Trp is calculated using the 6-311G** basis set (see text). $E(Na^+)$ (HF/6-31G**) = −161.6592883 Hartrees. $E(Na^+)$ (HF/6-311G**) = −161.6642329 Hartrees.

2.7 Appendix 1 – Details of Calculations

All calculations in this chapter were performed using Gaussian 94, Rev. D3. [49], or Spartan (versions 4 and 5) [50]. Na^+ binding energies were determined by the supermolecule method in which the total energy of the complex was minimized, and minimum energies of the individual partners were subtracted. In nearly all cases, Hartree-Fock theory was used with the 6-31G** basis set. The total energies of the individual molecules and their Na^+ complexes are given in Table 2.4 and Table 2.5.

Because bromine is not defined in the 6-31G** basis set, the calculations for 5-Br-Trp had to be performed using the 6-311G** basis set. The binding energy from

Aromatic	6-31G** minimum energy of Na^+ complex with 5-membered ring (Hartrees)
Trp	–523.1884613
5-F-Trp	–622.0314230
5,7-F_2-Trp	–720.8730516
5,6,7-F_3-Trp	–819.7020001
4,5,6,7-F_4-Trp	–918.5359377
7-aza-Trp	–539.1810810
5-CN-Trp	–614.9096883
Bf-Ala	–542.9939364

Table 2.5: Total energies of the 5-ring Na^+ complexes.

the 6-311G** calculation was extrapolated back to a 6-31G** binding energy based on a linear correlation of Na^+ binding energies calculated in both basis sets as shown in Figure 2.12. Clearly, Na^+ binding energies calculated using the two basis sets correlate well, although binding energies calculated with the larger 6-311G** basis set are a constant 2.25 kcal/mol less than those calculated with the 6-31G** basis set.

In addition to the total energies of the monomers, the electrostatic potential surfaces of each molecule were calculated at the HF/6-31G** level (HF/6-311G** for 5-Br-Trp). These are presented in Figure 2.13. All are displayed on an equal energetic scale, ranging from –25 kcal/mol (red) to +25 kcal/mol (blue).

2.8 Appendix 2 – Related Cation-π Calculations

In addition to the calculations previously described in this chapter, several related calculations involving simple ions binding to aromatics were performed and are briefly presented in this Appendix.

To evaluate the effects of both theoretical method and basis set size on the calculated $Na^+ \cdots$ benzene binding energies, a series of calculations was performed (Table 2.6). All complexes were constrained to have C_{6v} symmetry and are fully geometry optimized. It is clearly seen that using Hartree-Fock theory, the binding energies

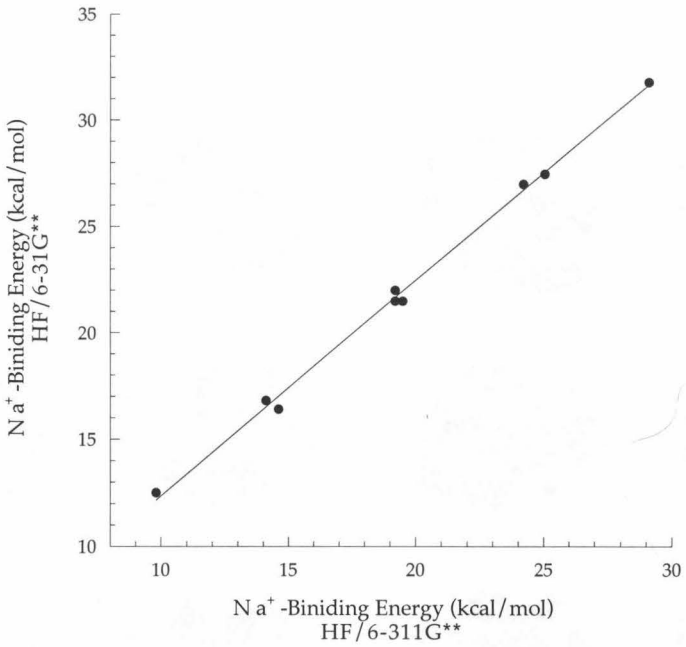


Figure 2.12: Plot of HF/6-311G** vs. HF/6-31G** Na^+ binding energies for a series of aromatics. The data fit the line $y = 2.25 + 1.01x$ with a correlation coefficient of $R = 0.998$. The aromatics, left to right, are: 1,3,5-trifluorobenzene, 1,4-difluorobenzene, 1,4-dichlorobenzene, chlorobenzene, fluorobenzene, 5-cyanoindole, benzene, 5-fluoroindole, and aniline.

decrease with increasing basis set size. Although consideration of basis set superposition error might mitigate some of this effect, there is nonetheless a substantial range of energies. Interestingly, both HF and B3LYP theory using the 6-31G** basis set perform quite well with respect to predicting the experimentally determined binding energies. This may be due to fortuitous cancellation of errors [46].

In addition to the calculations involving Na^+, a series of calculations involving ammonium ion (NH_4^+) binding to benzene was performed. In addition to the *ab initio* calculations shown in Table 2.7, calculations using the AM1 and PM3 semi-empirical methods are presented. These methods perform poorly, and both severely underestimate the experimentally-determined binding energy of the complex. Similar to the calculations for the $Na^+ \cdots$ benzene complex, most methods and basis sets underestimate the experimental binding energy. This is likely due to the fact that Hartree-Fock methods neglect the effects of electron correlation, and thus underestimate some of the van der Waals attraction between the partners. Density functional

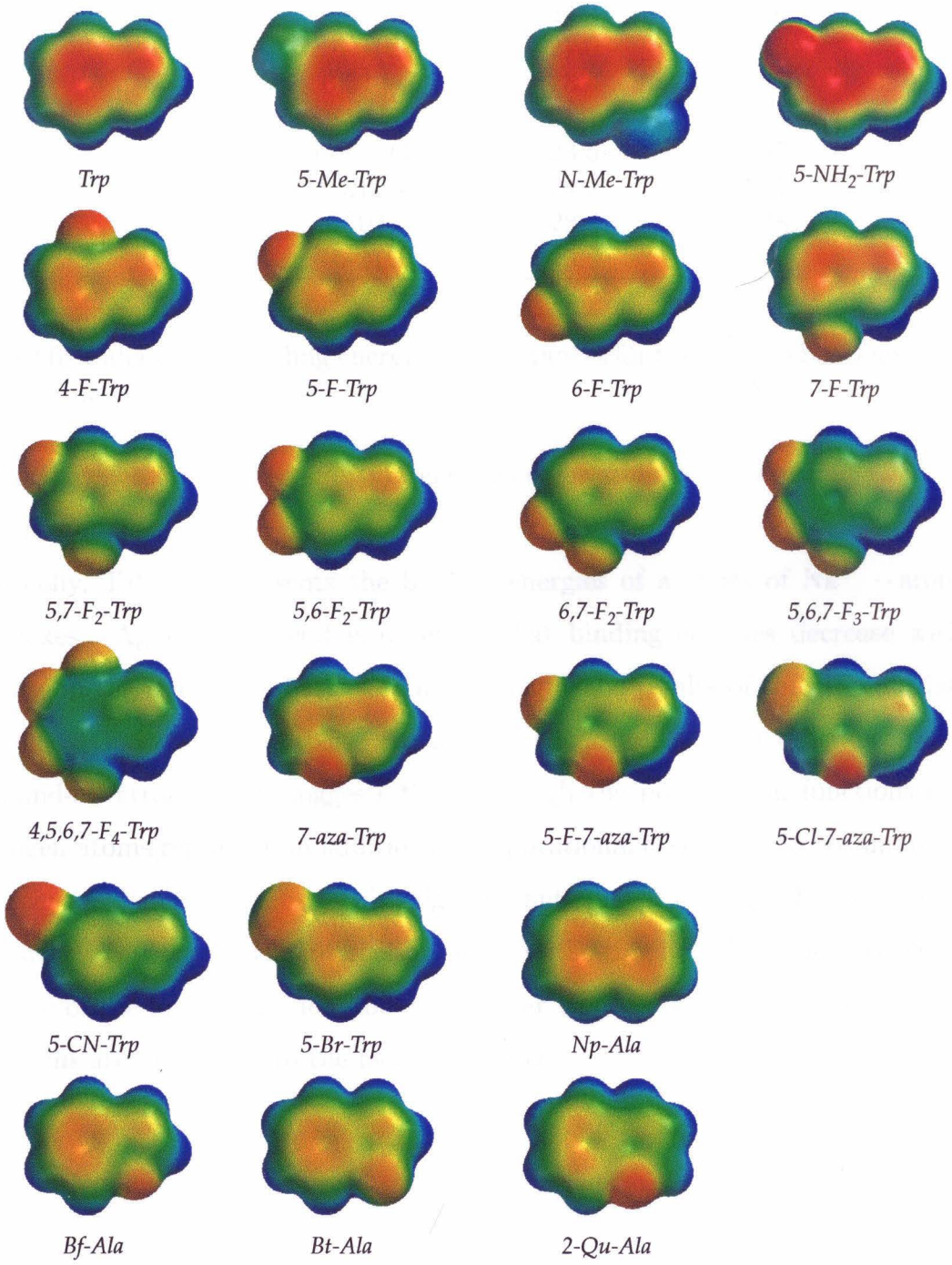


Figure 2.13: Calculated HF/6-31G** electrostatic potential surfaces for all aromatic side chains. *(Scale: –25 to +25 kcal/mol).*

Method	Basis Set	–Binding Energy (kcal/mol)	Distance Å
HF	3-21G*	34.6	2.36
HF	3-21G**	34.2	2.36
HF	6-31G*	27.0	2.45
HF	6-31+G*	23.5	2.48
HF	6-31G**	27.1	2.45
HF	6-31+G**	23.5	2.47
HF	6-311G**	24.2	2.47
B3LYP	6-31G**	28.4	2.38
BPW91	6-31G*	25.9	2.40
Experiment		28.0	

Table 2.6: Calculated –binding energies and distances for the $Na^+ \cdots$ benzene complex at several levels of theory.

methods, which include the effects of electron correlation, fare somewhat better than the HF methods.

Finally, Table 2.8 presents the binding energies of a series of $Na^+ \cdots$ aromatic complexes. Again, the trend is observed that binding energies decrease with an increase in the size of the basis set. Interestingly, the results of the HF calculations at the 6-31G* and 6-31G** basis sets are nearly identical – most differences are due to round-off errors. This suggests that although the polarization functions on the hydrogen atoms represent an additional computational cost, they do not improve the accuracy of the calculations. This might be anticipated because the hydrogens are not directly involved in the interaction between the ring and the cation. However, in the case of the ammonium ion above, a similar trend is seen, even though the NH_4^+ hydrogens are pointing into the face of the ring.

Method	Basis Set	Symmetry	–Binding Energy (kcal/mol)
AM1		C_{3v}	9.2
PM3		C_{3v}	6.8
HF	6-31G*	C_{3v}	14.1
HF	6-31G*	C_{2v}	15.3
HF	6-31G**	C_{3v}	14.1
HF	6-31G**	C_{2v}	15.3
B3LYP	6-31G**	C_{3v}	15.7
B3LYP	6-31G**	C_{2v}	17.5
B3LYP	MIDI!(6D)	C_{2v}	18.4
BPW-91	6-31G*	C_{3v}	14.5
BPW-91	6-31G*	C_{2v}	16.6
BPW-91	MIDI!(6D)	C_{2v}	17.5
Experiment			19.3

Table 2.7: Calculated –binding energies for the $NH_4^+ \cdots$ benzene complex at differing symmetries and levels of theory.

	Hartree-Fock				B3LYP
	3-21G**	6-31G*	6-31G**	6-311G**	6-31G**
Benzene	34.2	27.0	27.1	24.2	28.4
Fluorobenzene	28.1	21.9	22.0	19.2	24.4
1,4-Difluorobenzene	22.1	16.7	16.8	14.1	20.2
1,3,5-Trifluorobenzene	16.7	12.4	12.4	9.8	16.8
Chlorobenzene	28.7	21.4	21.5	19.2	23.5
1,4-Dichlorobenzene	23.7	16.4	16.4	14.6	19.1
Aniline	40.0	31.7	31.8	29.1	n.d.

Table 2.8: Calculated –binding energies for various $Na^+ \cdots$ aromatic complexes. All energies are in kcal/mol.

Bibliography

[1] A. Devillers-Thiéry, J. L. Galzi, J. L. Eiselé, S. Bertrand, D. Bertrand, and J. P. Changeux. Functional architecture of the nicotinic acetylcholine receptor: A prototype of ligand-gated ion channels. *J. Membrane Biol.*, 136:97–112, 1993.

[2] A. Karlin and M. H. Akabas. Toward a structural basis for the function of nicotinic acetylcholine receptors and their cousins. *Neuron*, 15:1231–1244, 1995.

[3] F. Hucho, V. I. Tsetlin, and J. Machold. The emerging three-dimensional structure of a receptor – the nicotinic acetylcholine receptor. *Eur. J. Biochem.*, 239:529–557, 1996.

[4] E. Kubalek, S. Ralston, J. Lindstrom, and N. Unwin. Location of subunits within the acetylcholine receptor by electron image analysis of tubular crystals from *Torpedo marmorata*. *J. Cell Biol.*, 105:9–18, 1987.

[5] N. Unwin. Nicotinic acetylcholine receptor at 9 Å resolution. *J. Mol. Biol.*, 229:1101–1124, 1993.

[6] N. Unwin. Projection structure of the nicotinic acetylcholine receptor: Distinct conformations of the α subunits. *J. Mol. Biol.*, 257:586–596, 1996.

[7] A. Miyazawa, Y. Fujiyoshi, M. Stowell, and N. Unwin. Nicotinic acetylcholine receptor at 4.6 Å resolution: Transverse tunnels in the channel wall. *J. Mol. Biol.*, 288:765–786, 1999.

[8] N. Unwin. Acetylcholine receptor channel imaged in the open state. *Nature*, 373:37–43, 1995.

[9] M. Noda, H. Takahashi, T. Tanabe, M. Toyosato, Y. Furutani, T. Hirose, M. Asai, S. Inayama, T. Miyata, and S. Numsa. Primary structure of α-subunit

precursor of *Torpedo californica* acetylcholine receptor deduced from cdna sequence. *Nature*, 299:793–797, 1982.

[10] M. Noda, H. Takahashi, T. Tanabe, M. Toyosato, S. Kikyotani, T. Hirose, M. Asai, H. Takashima, S. Inayama, T. Miyata, and S. Numa. Primary structures of β and δ subunit precursors of *Torpedo californica* acetylcholine receptor deduced from cdna sequences. *Nature*, 301:251–255, 1983.

[11] M. Noda, H. Takahashi, T. Tanabe, M. Toyosato, S. Kikyotani, Y. Furutani, T. Hirose, H. Takashima, S. Inayama, T. Miyata, and S. Numa. Structural homology of *Torpedo californica* acetylcholine receptor subunits. *Nature*, 302:528–532, 1983.

[12] M. Noda, Y. Furutani, H. Takahasghi, M. Toyosato, T. Tanabe, S. Shimizu, S. Kikyotani, T. Kayano, T. Hirose, S. Inayama, and S. Numa. Cloning and sequence analysis of calf cdna and human genomic dna encoding α-subunit precursor of muscle acetylcholine receptor. *Nature*, 305:818–823, 1983.

[13] T. Claudio, M. Ballivet, J. Patrick, and S. Heinemann. Nucleotide and deduced amino acid sequences of *Torpedo californica* acetylcholine receptor γ subunit. *Proc. Natl. Acad. Sci. USA*, 80:1111, 1983.

[14] M. Mishina, T. Kurosaki, T. Tobimatsu, Y. Morimoto, M. Noda, T. Yamamoto, M. Terao, J. Lindstrom, T. Takahashi, M. Kuno, and S. Numa. Expression of functional acetylcholine receptor from cloned cDNAs. *Nature*, 307:604–608, 1984.

[15] P. Charnet, C. Labarca, R. J. Leonard, N. J. Vogelaar, L. Czyzyk, A. Gouin, N. Davidson, and H. A. Lester. An open-channel blocker interacts with adjacent turns of α-helices in the nicotinic acetylcholine receptor. *Neuron*, 2:87–95, 1990.

[16] M.H. Akabas, D.A. Stauffer, M. Xu, and A. Karlin. Acetylcholine receptor channel structure probed in cysteine-substitution mutants. *Science*, 258:307–310, 1992.

[17] H. Zhang and A. Karlin. Identification of acetylcholine receptor channel-lining residues in the m1 segment of the β-subunit. *Biochemistry*, 36:15856–15864, 1997.

[18] H. Zhang and A. Karlin. Contribution of the β subunit m2 segment to the ion-conducting pathway of the acetylcholine receptor. *Biochemistry*, 37:7952–7964, 1998.

[19] P. Blount and J.P. Merlie. *Neuron*, 3:349–357, 1989.

[20] S. E. Pedersen and J. B. Cohen. *d*-tubocurarine binding-sites are located at α-γ and α- δ subunit interfaces of the nicotinic acetylcholine-receptor. *Proc. Natl. Acad. Sci. USA*, 87:2785–2789, 1990.

[21] C. Czajkowski and A. Karlin. Agonist binding site of *Torpedo* electric tissue nicotinic acetylcholine receptor. *J. Biol. Chem.*, 266:22603–22612, 1991.

[22] R. E. Middleton and J. B. Cohen. Mapping of the acetylcholine binding site of the nicotinic acetylcholine receptor: [^{3}H]nicotine as an agonist photoaffinity label. *Biochemistry*, 30:6987–6997, 1991.

[23] J.-L. Galzi and J.-P. Changeux. Neuronal nicotinic receptors: Molecular organization and regulations. *Neuropharmacology*, 34:563–582, 1995.

[24] M. Dennis, J. Giraudat, F. Kotzyba-Hibert, M. Goeldner, C. Hirth, J.-Y. Chang, C. Lazure, M. Chrétien, and J.-P. Changeux. Amino acids of the *Torpedo marmorata* acetylcholine receptor α subunit labeled by a photoaffinity ligand for the acetylcholine binding site. *Biochemistry*, 27:2346–2357, 1988.

[25] J.L. Galzi, F. Revah, D. Black, M. Goeldner, C. Hirth, and J.-P. Changeux. Identification of a novel amino acid α-tyrosine 93 within the cholinergic ligands-binding sites of the acetylcholine receptor by photoaffinity labeling. *J. Biol. Chem.*, 265:10430–10437, 1990.

[26] P. N. Kao, A. J. Dwork, R. R. J. Kaldany, M. L. Silver, J. Wideman, S. Stein, and A. Karlin. Identification of the α subunit half-cystine specifically labeled by an affinity reagent for the acetylcholine receptor binding site. *J. Biol. Chem.*, 259:11662–11665, 1984.

[27] D. C. Chiara and J. B. Cohen. Identification of amino-acids contributing to high and low affinity. *d*-tubocurarine (dtc) sites on the *Torpedo* nicotinic acetylcholine-receptor (nAChR) subunits. *FASEB J.*, 6:A106–A106, 1992.

[28] S.N. Abramson, W. Fenical, and P. Taylor. Lophotoxins: Irreversible active-site-directed inhibitors of nicotinic acetylcholine receptors. *Drug Develop. Res.*, 24:297–312, 1991.

[29] J.B. Cohen, S.D. Sharp, and Liu W.S. Structure of the agonist-binding site of the nicotinic acetylcholine receptor. *J. Biol. Chem.*, 266:23354–23364, 1991.

[30] D.A. Dougherty and D.A. Stauffer. Acetylcholine binding by a synthetic receptor. Implications for biological recognition. *Science*, 250:1558–1560, 1990.

[31] C. J. Noren, S. J. Anthony-Cahill, M.C. Griffith, and P.G. Schultz. A general method for site-specific incorporation of unnatural amino acids into proteins. *Science*, 244:182–188, 1989.

[32] J. Ellman, D. Mendel, S. Anthony-Cahill, C. J. Noren, and P. G. Schultz. Biosynthetic method for introducing unnatural amino-acids site-specifically into proteins. *Meth. Enzym.*, 202:301–336, 1991.

[33] J. D. Bain, C. G. Glabe, T. A. Dix, and A. R. Chamberlin. Biosynthetic site-specific incorporation of a non-natural amino acid into a polypeptide. *J. Am. Chem. Soc.*, 111:8013–8014, 1989.

[34] J. D. Bain, D. A. Wacker, E. E. Kuo, M. H. Lyttle, and A. R. Chamberlain. Preparation of chemically misacylated semisynthetic nonsense suppressor trnas employed in biosynthetic incorporation of non-natural residues into proteins. *J. Org. Chem.*, 56:4615–4625, 1991.

[35] J. D. Bain, C. Switzer, A. R. Chamberlin, and S. A. Benner. Ribosome-mediated incorporation of a non-standard amino acid into a peptide through expansion of the genetic code. *Nature*, 356:537–539, 1992.

[36] T. G. Heckler, L-H. Chang, Y. Zama, T. Naka, M. S. Chorghade, and S. M. Hecht. T4 RNA ligase mediated preparation of novel "chemically misacylate tRNA-Phes. *Biochemistry*, 23:1468–1473, 1984.

[37] M. W. Nowak, P. C. Kearney, J. R. Sampson, M. E. Saks, C. G. Labarca, S. K. Silverman, W. Zhong, J. Thorson, J. N. Abelson, N. Davidson, P. G. Schultz, D. A. Dougherty, and H. A. Lester. Nicotinic receptor binding site probed with unnatural amino-acid incorporation in intact cells. *Science*, 268:439–442, 1995.

[38] M. E. Saks, J. R. Sampson, M. W. Nowak, P. C. Kearney, F. Du, J. N. Abelson, H. A. Lester, and D. A. Dougherty. An engineered tetrahymena $tRNA^{Gln}$ for *in vivo* incorporation of unnatural amino acids into proteins by nonsense suppression. *J. Biol. Chem.*, 271:23169–23175, 1996.

[39] M. W. Nowak, J. P. Gallivan, S. K. Silverman, C. G. Labarca, D. A. Dougherty, and H. A. Lester. *In vivo* incorporation of unnatural amino acids into ion channels in a *Xenopus* oocyte expression system. *Meth. Enzymol.*, 293:504–529, 1998.

[40] M. D. Martin and A. Karlin. Functional effects on the acetylcholine receptor of mutiple mutations of γAsp174 and δAsp180. *Biochemistry*, 36:10742–10750, 1997.

[41] S. Mecozzi, A. P. West Jr., and D. A. Dougherty. Cation-π interactions in simple aromatics. Electrostatics provide a predictive tool. *J. Am. Chem. Soc.*, 118:2307–2308, 1996.

[42] S. Mecozzi, A. P. West Jr., and D. A. Dougherty. Cation-π interactions in aromatics of biological and medicinal interest: Electrostatic potential surfaces as a useful qualitative guide. *Proc. Natl. Acad. Sci. USA*, 93:10566–10571, 1996.

[43] M.E. O'Leary and M.M. White. Mutational analysis of ligand-induced activation of the *Torpedo* acetylcholine receptor. *J. Biol. Chem.*, 267:8360–8365, 1992.

[44] P.-J. Corringer, J.-L. Galzi, J.-L. Eiselé, S. Bertrand, J.-P. Changeux, and D. Bertrand. Identification of a new component of the agonist binding site of the nicotinic $\alpha 7$ homooligomeric receptor. *J. Biol. Chem.*, 270:11749–11752, 1995.

[45] Y. Fraenkel, J. M. Gershoni, and G. Navon. Acetylcholine interactions with tryptophan-184 of the α-subunit of the nicotinic acetylcholine receptor revealed by transferred Nuclear Overhauser Effect. *FEBS Lett.*, 291:225–228, 1991.

[46] J. C. Ma and D. A. Dougherty. The cation-π interaction. *Chem. Rev.*, 97:1303–1324, 1997.

[47] S. F. Boys and F. Bernardi. *Mol. Phys.*, 19:533–566, 1970.

[48] W. Zhong. *Physical Organic Chemistry on the Nicotinic Acetylcholine Receptor.* Ph.D. Thesis, California Institute of Technology, 1998.

[49] M. J. Frisch, G. W. Trucks, H. B. Schlegel, P. M. W. Gill, B. G. Johnson, M. A. Robb, J. R. Cheeseman, T. A. Keith, G. A. Petersson, J. A. Montgomery, K. Raghavachari, M. A. Al-Laham, V. G. Zakrzewski, J. V. Ortiz, J. B. Foresman, C. Y. Peng, P. A. Ayala, M. W. Wong, J. L. Andres, E. S. Replogle, R. Gomperts, R. L. Martin, D. J. Fox, J. S. Binkley, D. J. Defrees, J. Baker, J. P. Stewart, M. Head-Gordon, C. Gonzalez, and J. A. Pople. *Gaussian 94 (Revision D.3).* Gaussian, Inc., 1995.

[50] Spartan, version 5.0., 1995.

Chapter 3 Cation-π Interactions in Structural Biology

3.1 Introduction

The three-dimensional structure of a protein is uniquely determined by a delicate balance of weak interactions. Hydrogen bonds, salt bridges, and the hydrophobic effect all play important roles in folding a protein and establishing its final structure. In addition to these extensively studied interactions, the cation-π interaction [1–3] is increasingly recognized as an important noncovalent binding interaction of relevance to structural biology. Theoretical and experimental studies have shown that cation-π interactions can be quite strong, both in the gas phase and in aqueous media. A wide array of biological structures has been shown to employ cation-π interactions, with perhaps the most compelling cases involving the binding of the quaternary ammonium group of acetylcholine by both acetylcholinesterase [4] and the nicotinic acetylcholine receptor [5].[1] These studies and others have emphasized the role of cation-π interactions in biological recognition, such as receptor-ligand and enzyme-substrate interactions. In this chapter, we attempt to answer a broader question – How does Nature use cation-π interactions *within* protein structures? In doing so, we present a detailed analysis of the extent and nature of cation-π interactions that are intrinsic to a protein's structure and are likely to contribute a protein's stability. We find that energetically significant cation-π interactions are common in proteins – a "typical" protein will contain several. In addition, we document some significant preferences for certain amino acid pairs as partners in a cation-π interaction.

Important early work indicated a role for cation-π interactions in protein structures. Following work by Levitt and Perutz [6–8] which suggested that aromatic rings could be involved in a hydrogen bonding-like interaction with amino groups, a pioneering study by Burley and Petsko identified the "amino-aromatic" interaction [9], showing that amino groups tend to be positioned near aromatic rings within proteins. No distinction was made between neutral and charged NH groups in that study. It is now appreciated that the interaction of a cationic group with an aromatic – a cation-π interaction – is more favorable energetically by a full order of magnitude

[1] This work is discussed in Chapter 2.

than an analogous interaction involving a neutral amine. For example, ammonium ion binds to benzene with a strength of 19 kcal/mol [10], whereas the binding interaction between ammonia and benzene is less than 2 kcal/mol [11]. As such, it is best to separate interactions involving the cation Lys and Arg from those involving the neutral, amide side chains of Asn and Gln. Subsequent studies by Thornton and co-workers [12–17] modified the Burley and Petsko analysis, especially with regard to the amino-aromatic hydrogen bond. Although Thornton and co-workers observed the same preference for NH-containing groups to be located near aromatic amino acids, they suggest that it is not an amino-aromatic hydrogen bond that drives the formation of such structures. In particular, the side chains of Asn and Gln tend to stack parallel to aromatic side chains. This geometry precludes hydrogen bonding to the aromatic ring, but allows the amino group to satisfy its hydrogen bonding potential by donating a hydrogen bond to either solvent or other parts of the protein. Concerning cationic residues, explicit studies of Arg interacting with aromatic residues have been reported by Flocco and Mowbray [18], and by Thornton [14], and other efforts to search the PDB for cation-π interactions between ligands and proteins have been reported [19, 20]. A recent study by Schultz has shown that an engineered cation-π interaction can contribute to protein stability with a magnitude comparable to that of a hydrogen bond [21].

As is typical of such work, previous searches of the protein database relied primarily on geometric definitions of side chain interactions, focusing on when a cationic side chain displayed a certain distance/angle relationship to an aromatic side chain. Because of the substantially different geometries of Lys vs. Arg and Trp vs. Phe/Tyr (Figure 3.1), such comparisons can be problematic. In addition, not all cation-aromatic *contacts* represent energetically favorable, cation-π interactions. Unlike ion pairs (e.g., Lys/Asp), for which any close contact will be energetically favorable, the interaction between a cation and an aromatic can be attractive (over the face) or repulsive (on the edge). The electrostatic potential surfaces of the aromatics, which control such distinctions, can be complex,[2] and so it is difficult to clearly distinguish

[2]See: Chapter 2, Figure 2.4.

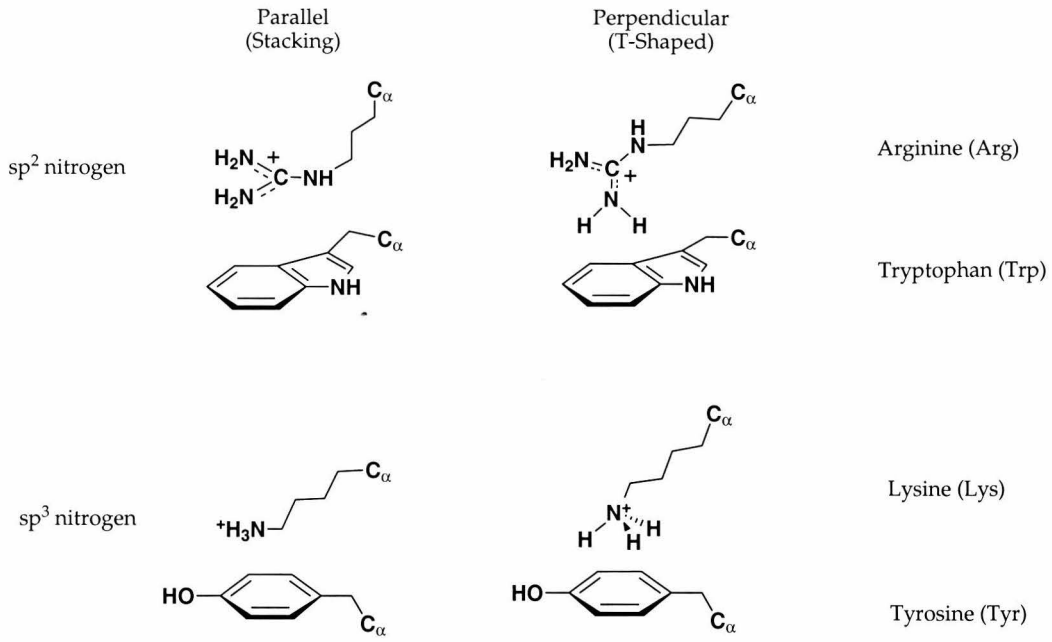


Figure 3.1: Limiting geometries for cation-π interactions. Top: Stacked and T-shaped alignments for arginine. Bottom: Lysine interacting with Phe, with the ε-carbon (left) or the NH_3^+ (right) making closest contact with the ring. These structures are not meant to indicate a preference for any particular pair or geometry.

attractive from repulsive cation-aromatic contacts using geometrical criteria (r, θ, ϕ) alone. For example, while the interaction of Lys with Phe can perhaps be adequately modeled considering only geometry - good interactions have the RNH_3^+ group "above" the phenyl ring - such arguments become much more challenging when the cation is Arg (T-shaped vs. stacked vs. intermediate) and the aromatic is Trp (6-ring vs. 5-ring). To circumvent this problem, and to put the diverse array of potential cation-π interactions on a more nearly equal footing, we have chosen to use *energy-based*, rather than geometry-based criteria in this study. Our goals in this study are two-fold. First, we wish to develop meaningful statistics for cation-π interactions for structures within the Protein Data Bank (PDB) [22]. Second, we wish to develop a simple, unambiguous protocol for identifying cation-π interactions which can be easily applied by other workers.

Within a protein chain, cation-π interactions can occur between the cationic side chains of either lysine (Lys, K) or arginine (Arg, R) and the aromatic side chains of phenylalanine (Phe, F), tyrosine (Tyr, Y), or tryptophan (Trp, W). Because histidine

can participate in cation-π interactions as either a cation (when the side chain is protonated), or as a π-system, and there can be some ambiguity on this issue, we do not consider histidine in this study. We assume Lys ($pK_a = 10.5$) is always protonated in this study.

There are several possible geometries for interactions involving arginine and lysine. With lysine, the obvious geometry places the NH_3^+ group over the ring. However the ε-carbon of Lys (CE) also carries a substantial positive charge, and putting it near an aromatic also constitutes a cation-π interaction. Arginine has two limiting geometries, a perpendicular, or T-shaped geometry that places the NH groups of the side chain in the face of the aromatic, and a parallel geometry in which the guanidinium group is stacked upon the face of the aromatic (Figure 3.1).

3.2 Methods

In this section we describe in some detail our strategy for identifying and ranking cation-π interactions in proteins. To provide an energetic evaluation of all potential cation-π interactions in a protein, we use a variant of the OPLS force field [23, 24]. Only cation-π interactions with binding interactions that rise above a certain energetic threshold are retained. The details of and justification for the protocol constitute the remainder of this section. Further discussion of this method appears in an appendix to this chapter.

Our development of a simple, general protocol to identify cation-π interactions within the PDB proceeded as follows:

1. Potential cation-π interactions from a test dataset of 68 proteins were identified using only geometric criteria.

2. Ab initio quantum mechanical calculations were performed on the candidates from the test dataset to evaluate their binding energies.

3. A force field–based method was developed to reproduce the trends in the ab initio data.

4. The force field–based method was then used to select energetically significant cation-π interactions from a larger dataset of 593 proteins.

To search the PDB, a computer program (CaPTURE - Cation-π Trends Using Realistic Electrostatics) was developed to calculate the distance between the cationic group (the ammonium nitrogen (NZ) in Lys or the guanidinium carbon (CZ) in Arg) and the centers of all aromatic rings (the centroids of the 6-membered rings in Phe, Tyr and Trp, as well as the 5-membered ring of Trp). We chose 68 proteins from the PDB whose resolution is better than 2.0 Å and which had been previously shown by Thornton's group to be non-homologous [16].

Using a 6.0 Å distance cutoff and no other geometrical constraints, 359 potential cation-π pairs were selected. Each pair was then reduced to a system that could be studied computationally. The cations Lys and Arg were represented as ammonium and guanidinium ions, and the aromatics, Phe, Tyr and Trp, were represented as benzene, phenol and indole, respectively. Using HF/6-31G** minimum energy structures for these fragments, we established the relative orientations of the partners using geometrical parameters from the PDB and converted each potential interaction to a Z-matrix format suitable for input to GAUSSIAN 94 [25]. The binding energy of each was determined using HF/6-31G** calculations with counterpoise corrections for basis set superposition error (BSSE). We appreciate that a higher level of theory (such as MP2) would provide a better estimate of the *gas phase* interaction energy for any pair [26, 27]. However, that is not our goal. We seek a simple criterion that will put all cation-aromatic contacts on a consistent scale, and it is clear that the *trends* from the level of theory applied here reproduce the trends of the higher levels. In addition, it must be remembered that even high resolution macromolecular structures have some structural ambiguity, and we feel that attempting very high level calculations on relatively "low level" geometries is not sensible.

Our ultimate goal is to develop a simple, rapid procedure to identify cation-π interactions, and a force field–based method is best suited to this. While standard force-field-methods are challenged to quantitatively model the cation-π interaction [28–30],

it is clear that force field–based methods can correctly reproduce trends in the binding energies. We therefore implemented within CaPTURE a subset of the OPLS force field [23] in which only electrostatic and van der Waals interactions are considered. We next calculated the OPLS binding energies for each of the 359 interactions described above, and compared them to the HF calculated binding energies. At first, the results were not encouraging, producing a poor correlation between the HF energies and the OPLS total energies (E_{tot}, which equals the sum of an electrostatic term, E_{es}, and a van der Waals term, E_{vdW}). Previous studies have shown that a large component of the cation-π interaction is electrostatic, and that trends in the electrostatic component determine trends in cation-π binding ability [31]. Thus, we attempted to correlate the electrostatic component of the OPLS binding energy (E_{es}) with the total ab initio binding energy. We find that these measurements correlate well for the different cation-π pairs (Figure 3.2), and represent an enormous computational savings. The correlations are better when Lys is the cation than when Arg is the cation. This is possibly because the larger, more complex Arg side chain has several different binding modes available, and these produce variations in van der Waals interactions, which are now excluded in our force field calculation, but are included in the ab initio calculations. In addition, the HF structures were built under the assumption that the δ carbon (CD) of Arg is in the same plane as the guanidinium unit. However, in some protein structures this is not so, and the OPLS calculations use the structure directly from the PDB file. This introduces some further scatter in the data. Our major concern, however, is with the *trends* in the data – we want to know when one cation-π interaction is stronger than another. We feel this approach is more than adequate for such purposes.

Because force field calculations are computationally inexpensive, we included the "methyl-groups" for the cationic side chains (Arg, CD; Lys, CE) on moving to the much larger dataset of proteins discussed below. Thus, Lys was now represented as methylammonium and Arg as methylguanidinium. This change was motivated by the fact that a significant fraction of the positive charge in Lys and Arg is associated with these carbons. Methyl groups were represented as united atoms because of

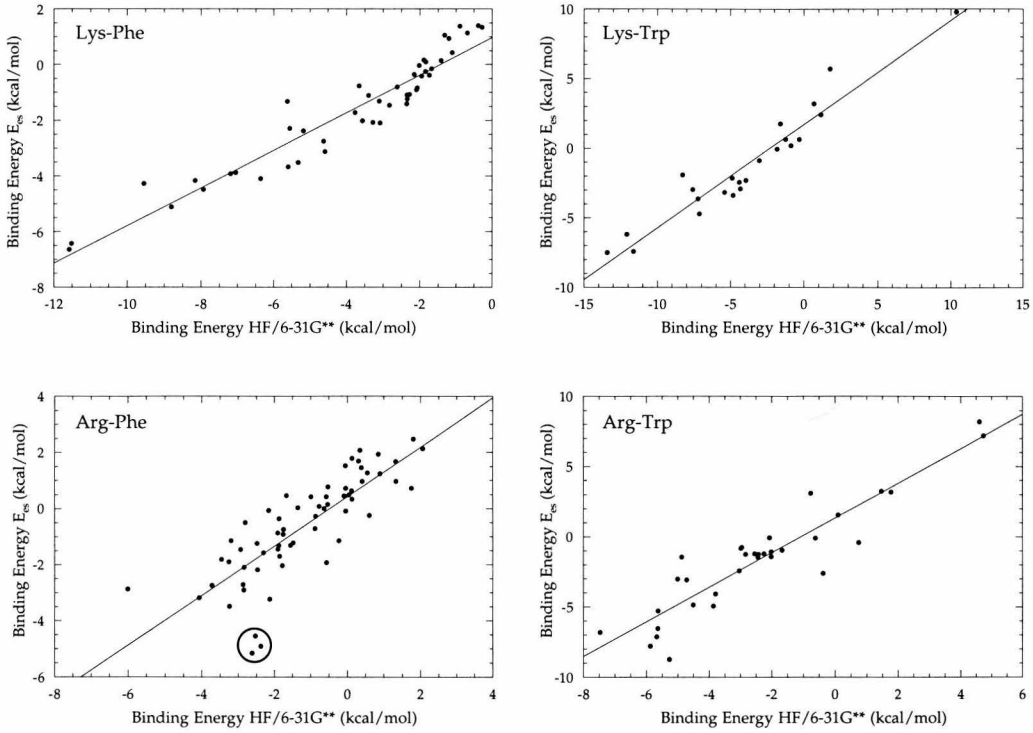


Figure 3.2: Plots of E_{es} vs. counterpoise-corrected, HF 6-31G** binding energies for selected cation-π interactions. The data fit to the following lines: Lys-Phe $y = 0.98 + 0.68x, R = 0.95$, Lys-Trp $y = 1.72 + 0.74x, R = 0.96$, Arg-Phe $y = 0.42 + 0.88x, R = 0.81$, Arg-Trp $y = 1.33 + 1.23x, R = 0.92$. Three outliers are circled in the Arg-Phe plot. Inspection of these pairs reveals that they contain spurious close contacts (see text), which lead to erroneously high energies in the HF calculations, but not in the OPLS calculations, because no van der Waals term is included in the latter. If they are removed from the plot, the correlation coefficient improves to 0.85. Note that due to complications concerning the representation of Tyr (see text), such plots are not meaningful when Tyr is the aromatic.

uncertainty in the position of the hydrogens. The inclusion of methyl groups on the aromatics does not significantly alter their cation-π binding ability [32], and they were not added. The hydrogens of the aromatics were placed in idealized geometries and treated explicitly (not as united atoms) using the standard parameters in the OPLS force field.

Tyrosine presents a special problem for this analysis. The position of the proton of the phenolic OH is typically not well-determined in macromolecular crystal structures, and its location can significantly affect cation binding ability through an interaction of the ion with the strong OH bond dipole. In principle, we could treat the OH of

tyrosine as a united atom, as we have treated other groups with uncertainty in the positions of the hydrogens. Unfortunately, this presents a new problem, as the partial negative charge on this atom strongly attracts cations, but not through a cation-π interaction. Since phenylalanine and tyrosine are nearly equivalent in their idealized cation-π binding ability [31, 32],[3] we have chosen to treat them with identical charge parameters. The phenolic oxygen of tyrosine retains its steric parameters, but has zero charge. To make the ring electrostatically equivalent to Phe, a dummy charge is placed in an equivalent location to that of the *para*-hydrogen of phenylalanine, and all other atomic charges are set equal to those of phenylalanine. Thus, in interactions of a cation with Tyr, we are considering *only* the cation-π interaction. In the actual protein, the cation-tyrosine interaction may be larger due to interactions with the oxygen (see below).

Another advantage of excluding E_{vdW} from our calculation is that many of the most favorable cation-π interactions we find in proteins produce spurious, *repulsive* E_{vdW} terms. This results from the inherently low resolution of macromolecular crystal structures, which occasionally have unrealistic atom-atom contacts. Such pairs would be rejected from a search for cation-π interactions if E_{tot} were the evaluation criterion. Since electrostatics are much less sensitive to close contacts ($1/r$) than van der Waals repulsions ($1/r^{12}$), the effects of such small geometry errors are minimized when only E_{es} is considered.

With this computationally efficient, energy-based criterion established, we could now examine a larger number of proteins (see below), and we could expand the interaction distance cutoff from 6 Å to 10 Å. Within this expanded search, we found some cation-π pairs involving tryptophan that gave a favorable value of E_{es} but had the cation quite far from the aromatic rings. Although these structures are certainly attractive in the gas phase, they are likely attenuated in the interior of a protein. Because of this, a simple, geometry-based criterion was added that would eliminate these long range structures. To do this, we asked whether a water-sized probe (a sphere

[3]The calculated cation-π binding energies and electrostatic potential surfaces of these residues can be found in Chapter 2.

2.8 Å in diameter) could fit between the van der Waals surfaces of the cation and the aromatic at their point of closest approach. If the water molecule fits, the structure is rejected, regardless of its electrostatic energy. In practice, this criterion eliminates a small number of distant interactions. Indeed, in some structures, inspection of the full protein rather than just the cation-π pair we had selected for evaluation, revealed intervening atoms from the protein separating the "interacting" partners. We felt it was sensible to reject such structures.

It remains to choose a threshold value for E_{es}, below which a pair is considered to experience a cation-π interaction. The simplest model would assume that if E_{es} is less than zero, i.e., if the electrostatic interaction is favorable, then the pair should be counted as a cation-π interaction. However, inspection of structures with E_{es} only slightly below zero shows that these interactions are much too weak to consider. They may be significant in the gas phase, but we consider them unlikely to contribute to protein stability. On the other hand, it is clear that if $E_{es} \leq$ -2.0 kcal/mol, then the pair is experiencing a significant cation-π interaction. Also, if $E_{es} >$ -1.0 kcal/mol, no cation-π interaction should be considered. For -2.0 $< E_{es} \leq$ -1.0 kcal/mol, the choice is not so clear. Some structures are clearly desirable, while in other cases the interacting partners are too far apart, even though there is no 2.8 Å gap. In order to distinguish these, we consider the van der Waals term of the OPLS interaction energy as an indicator of whether the pair is interacting significantly. It is safe to include E_{vdW} for these relatively weaker interactions, because they are generally not the closest contacts, and therefore not susceptible to spurious van der Waals interactions. We conclude that if E_{vdW} is $\leq$ -1.0 kcal/mol, the interaction is significant. We emphasize that these final refinements to the protocol are meant to produce the best possible selection criterion, but they do not have a major impact on the final list. None of the global conclusions presented below would be substantially altered if they were not included.

To summarize, our protocol for selecting cation-π interactions is as follows: All cation-π pairs (K or R with F, Y, or W) within 10 Å of each other are considered. If there is a gap large enough to insert a water molecule (2.8 Å separation between

van der Waals surfaces at closest contact), the structure is rejected, and the residues are considered non-interacting. For the remaining interacting pairs, the OPLS electrostatic energy, E_{es}, is evaluated. If $E_{es} \leq$ -2.0 kcal/mol, the pair is counted as a cation-π interaction. If $E_{es} >$ -1.0 kcal/mol, the structure is rejected. If -2.0 $< E_{es} \leq$ -1.0 kcal/mol, the structure is retained only if E_{vdW} is $\leq$ -1.0 kcal/mol. It is worth remembering that the interaction energies we will discuss below are only the OPLS electrostatic energies. The actual interaction energy – the true magnitude of the cation-π interaction – is larger by an amount equal to the van der Waals interaction energy. For most pairs considered, E_{vdW} is comparable to E_{es}, and so the true cation-π interaction energy is roughly twice as large as E_{es}.

3.3 Results and Discussion

Using the above criteria, we scanned a larger dataset of representative protein crystal structures, taken from the "PDB Select" list of Hobohm and Sander [33, 34]. We initially divided the proteins into two sets. The first consisted of 323 proteins that exist as a single subunit (Table 3.1); the second consisted of 270 proteins that contained multiple subunits (Table 3.2).

In most analyses, however, no significant differences between these two sets were found, and thus the combined set of 593 proteins is considered unless otherwise noted. All proteins considered had resolutions better than 2.5 Å. For evaluating possible cation-π interactions, residues with fractional occupancies < 0.95 were rejected, and structures that contained residue insertions were not considered. For the combined sets, 230,504 residues were considered, producing 14,030 interacting pairs (structures within 10 Å of each other and containing no "gap", Table 3.3), and 2,994 significant cation-π interactions (Table 3.4).

Figure 3.3 gives a visual representation of our selection procedure. All Lys/Phe pairs from single subunit proteins within the 10 Å cutoff are shown. They are divided into three categories: those that are rejected based on the gap criterion (open circles); those that are interacting (no gap) but have $E_{es} >$ -1.0 kcal/mol (blue circles); and

119L	1AM3	1CFB	1GCB	1LBU	1PII	1TCA	1YER	2PII
153L	1AMM	1CFR	1GKY	1LCL	1PLC	1TDF	1YGS	2PLC
1A0P	1AMP	1CHD	1GND	1LIS	1PMI	1TDX	1YTW	2POR
1A17	1AMX	1CLC	1GOH	1LIT	1PNE	1TFE	1ZIN	2PTH
1A1X	1AN8	1CNV	1GPC	1LKI	1POA	1TFR	1ZXQ	2RN2
1A26	1ANF	1CPO	1GSA	1LML	1POC	1THV	2A0B	2SAK
1A62	1AOL	1CSH	1GVP	1MAI	1POT	1THX	2ABK	2SIL
1A68	1AOP	1CSN	1HA1	1MAZ	1PPN	1TIB	2ACY	2SN3
1A6Q	1AQB	1CTJ	1HFC	1MBD	1PPT	1TIF	2AYH	2SNS
1A7J	1ARB	1CYX	1HJP	1MML	1PTA	1TML	2BAA	2STV
1A8E	1ARV	1DAD	1HLB	1MRJ	1PTQ	1TSP	2BGU	2TCT
1A8I	1ASH	1DAR	1HMY	1MRP	1PTY	1TUL	2CAE	2TGI
1A9S	1AT0	1DHR	1HOE	1MSC	1PUD	1TYS	2CBA	2WEA
1AA0	1AUA	1DHS	1HTP	1MSI	1QNF	1UAE	2CTC	3B5C
1AAC	1AUK	1DHY	1HXN	1MSK	1QUF	1UBI	2CYP	3CHY
1ABA	1AWD	1DUN	1IDK	1MUP	1R69	1UBY	2DKB	3CLA
1ACC	1AXN	1DXY	1IDO	1MZM	1RA9	1UCH	2DPG	3COX
1AD2	1AYL	1ECA	1IGD	1NAR	1RCF	1UTG	2DRI	3CYR
1ADS	1AZO	1ECL	1IHP	1NEU	1REC	1UXY	2END	3GRS
1AF7	1BA1	1EDE	1INP	1NFN	1RHS	1V39	2ENG	3LZT
1AH7	1BDO	1EDG	1IRK	1NIF	1RIE	1VCC	2ERL	3NLL
1AHO	1BEO	1EDT	1ISO	1NLS	1RLW	1VHH	2FDN	3NUL
1AIE	1BFD	1ERV	1ITG	1NOX	1RMD	1VID	2FHA	3PTE
1AIL	1BFG	1ESC	1IXH	1NPK	1RMG	1VIF	2GDM	3SEB
1AJ2	1BGC	1EUU	1JDW	1NXB	1RNL	1VIN	2HBG	3TSS
1AJ6	1BGP	1EZM	1JER	1OIS	1RPO	1VJS	2HFT	3VUB
1AJJ	1BKF	1FDR	1JPC	1ONC	1RRO	1VLS	2HTS	4MT2
1AK0	1BP1	1FDS	1KID	1OPD	1RSS	1VNC	2I1B	4XIS
1AK1	1BRT	1FIT	1KIT	1OPR	1RYT	1VSD	2ILK	5EAU
1AK5	1BTN	1FKX	1KNB	1OPY	1SBP	1WAB	2LBD	5ICB
1AKO	1BV1	1FNA	1KPF	1OYC	1SFE	1WBA	2LIV	5P21
1AKZ	1BYB	1FTS	1KTE	1PDA	1SFP	1WER	2MCM	5PTI
1AL3	1C52	1FUA	1KUH	1PEA	1SKZ	1WHI	2OMF	6CEL
1ALO	1CEM	1FUS	1KVU	1PGS	1SMD	1WHO	2PGD	7RSA
1ALU	1CEO	1G3P	1LAM	1PHE	1SRA	1XJO	2PHY	8ABP
1ALY	1CEX	1GAI	1LBA	1PHP	1SVB	1XNB	2PIA	

Table 3.1: PDB ID Codes for single-subunit proteins.

1A1I	1AOZ	1BTK	1FJM	1IPS	1MSP	1SFT	1WSY	2PHL
1A28	1APY	1CEW	1FMT	1IPW	1MUC	1SLT	1XBR	2POL
1A2P	1AQ0	1CFY	1FT1	1ISU	1MUG	1SMN	1XGS	2PSP
1A2Y	1AQ6	1CHK	1FTP	1ITB	1NBA	1SMP	1XIK	2RSL
1A34	1AQZ	1CHM	1FUR	1JAC	1NBC	1SMT	1XVA	2RSP
1A6J	1ASK	1CKA	1FVK	1JET	1NCI	1SPU	1XYZ	2SCP
1A74	1ATL	1CKN	1FWC	1JFR	1NOZ	1STM	1YAI	2SIC
1A7T	1ATZ	1CNT	1GAR	1JHG	1NP1	1SVP	1YAS	2SPC
1AA7	1AU1	1CPC	1GC1	1JLY	1NSG	1TAD	1YCQ	2TMK
1ABR	1AUR	1CSB	1GDO	1JMC	1NUL	1TAF	1YCS	2TRC
1ADO	1AUV	1CSE	1GGG	1JSU	1NWP	1TC3	1YPD	2TYS
1AER	1AVM	1CYD	1GIF	1JXP	1ONR	1TDT	1YRN	2VHB
1AFR	1AW8	1DEA	1GNW	1KIM	1OSP	1THT	1YTB	2VIU
1AFW	1AWC	1DEL	1GOT	1KNY	1OTG	1TII	1YTF	3DAA
1AGJ	1AWS	1DFJ	1GPM	1KPT	1PBO	1TRK	1YVE	3DAP
1AGQ	1AX4	1DJX	1GTM	1KVE	1PCF	1TUP	2ARC	3MDD
1AIJ	1AXI	1DKZ	1GTQ	1KWA	1PDN	1TVX	2BBK	3MIN
1AIK	1AYM	1DOK	1GUQ	1KZU	1PHN	1UMU	2BOP	3PCG
1AJS	1AYY	1DOR	1HAV	1LAT	1PNK	1UNK	2CCY	3SDH
1AK4	1AZS	1DOS	1HCR	1LGH	1POI	1URN	2CHS	3ULL
1ALV	1BBP	1DUB	1HGX	1LKK	1PRC	1VCA	2DYN	4AAH
1AM7	1BDM	1DUP	1HJR	1LMB	1PRX	1VDE	2FIV	4CPA
1AMU	1BEB	1E2A	1HSB	1LT5	1PYA	1VDF	2HHM	4PGA
1AN9	1BEN	1ECE	1HSS	1LTS	1QTQ	1VHR	2HPD	4PGM
1AO6	1BFT	1ECP	1HTR	1LUC	1REG	1VMO	2KIN	5CSM
1AOC	1BHM	1EDM	1HUL	1MKA	1REQ	1VPF	2MAS	5HPG
1AOH	1BKR	1EFV	1IDA	1MLD	1RGE	1VPS	2MPR	6GSV
1AOJ	1BNC	1ETP	1IGN	1MNM	1RHO	1VWL	2MSB	7AHL
1AOQ	1BND	1EXN	1IHF	1MOL	1RLA	1WDC	2MTA	8ATC
1AOR	1BOV	1FIE	1IIB	1MPG	1RVA	1WPO	2NAC	8RXN

Table 3.2: PDB ID Codes for multi-subunit proteins.

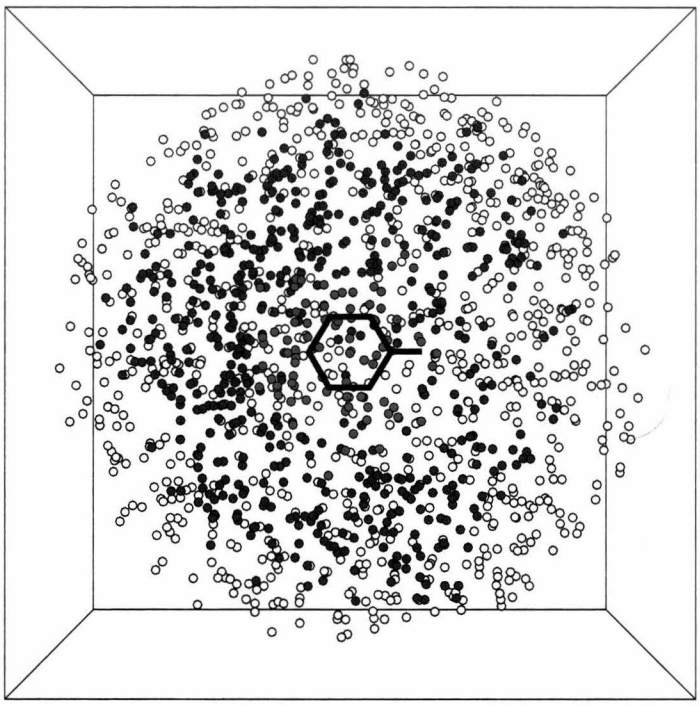


Figure 3.3: Scatter plot for all Lys/Phe pairs from all 323 single subunit proteins. The circle denotes the location of the side chain N. Each side of the cube is 10 Å. The phenyl ring plus the β carbon are denoted by black lines. The color code is as follows: red – a cation-π interaction; blue – an interacting pair; white – structures rejected by the gap criterion. In this projection view, a few open/blue circles are seen to lie over the ring, but they are too far away to have a favorable cation-π interaction.

cation-π interactions (red circles). Clearly our selection is "sensible" – the three classes form concentric rings about the centroid of the aromatic ring.

It is worth noting here a significant advantage of energy-based selection criteria over geometry-based methods. Like other workers, our initial geometric screen involved a distance (r) from the center of positive charge (NZ of Lys or CZ of Arg) to the aromatic, but there is a danger in rejecting structures solely on the basis of r. For example, there are many structures in which the NZ of Lys will be more than 5 Å from the aromatic, but the CE will be well within 5 Å. Since the CE of Lys has a substantial positive charge, such a structure should be considered as a possible cation-π interaction, and our calculations reflect this. An example of such a structure can been found in Section 3.4. While we considered all pairs with r $\leq$ 10 Å, in 99% of our accepted cation-π structures, r is $\leq$ 6 Å when both NZ and CE of Lys or CZ

and CD of Arg are considered, while 88% have r $\leq$ 5 Å.

With 2,994 cation-π interactions in 230,504 residues, there is an average of 1 energetically significant cation-π interaction for every 77 residues in a protein. This number does not vary substantially with the length of the protein, although there is certainly some scatter. For example, the 126 amino acid mutant form of the protein human fibroblast growth factor (PDB: 1BFG) contains 6 significant cation-π interactions, while the 323 amino acid protein penicillopepsin (PDB: 2WEA) is the largest single chain protein studied that contains no energetically significant cation-π interactions, although this protein contains only 5 lysines and no arginines. The number of cation-π interactions per residue is the same whether single chain or multi-subunit proteins are considered.

For comparison, one might ask how common salt bridges are in proteins. The energetic significance of salt bridges continues to be an issue of debate – a debate that we will consider further in Chapter 4. Nevertheless, to get a sense of the relative frequency of the two types of interactions, we considered all ion pairs (Lys/Asp, Lys/Glu, Arg/Asp, Arg/Glu) that meet the "interacting pair" criterion, i.e., they have no gap. We find that salt bridges are quite common, with almost 27,000 occurring in our collection of proteins (compared to 14,030 cation-aromatic interacting pairs). There is no significant preference for one pair over another.

The most common cation-π interaction is between neighboring residues in the sequence, with 7.3% of the interactions occurring between adjacent residues (Figure 3.4). Interactions between residues at the i and $i+4$ positions are the second most common, suggesting that cation-π interactions may commonly occur within α-helices, as in the structure of the vaccinia virus protein VP39 (PDB: 1V39) shown in Figure 3.5.

Table 3.3 summarizes the likelihood that individual amino acids will be involved in an interacting pair, while Table 3.4 lists analogous data for cation-π interactions. Table 3.4 presents two sets of numbers: all cation-π interactions and those with E_{es} $\leq$ -4.0 kcal/mol. The latter set contains strong interactions that, by any criterion, would be considered to make a substantial contribution to protein stability. Several

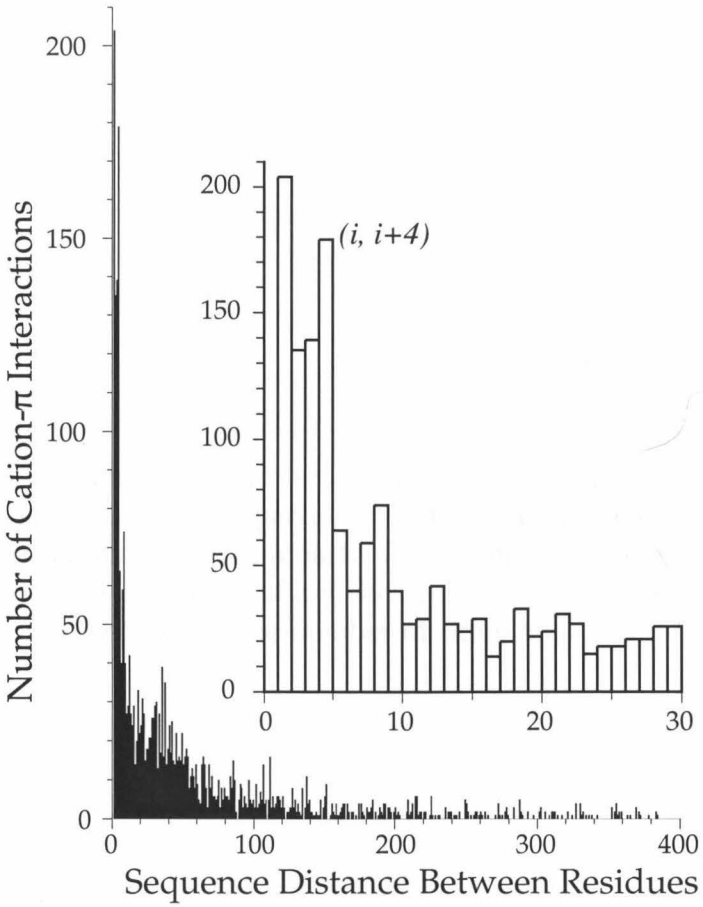


Figure 3.4: Plot of the number of cation-π interactions vs. the sequence distance of the interacting pair. The inset highlights interactions between the i and $i+4$ residues that may be located on α-helices. Only single chain proteins were considered for this analysis.

interesting trends concerning individual amino acids are evident.

3.3.1 Cationic Amino Acids

With respect to the cationic amino acids, Table 3.3 indicates a striking preference for the side chain of arginine to be located near aromatic side chains in proteins. Over 70% of all Arg side chains are near an aromatic side chain, consistent with earlier studies of the arginine-aromatic interaction [12]. In Table 3.4, which considers only favorable structures, we again note a strong preference for arginine over lysine that mimics the trends in Table 3.3.

It is perhaps surprising that arginine is more likely to be found in a cation-π

Amino acid	Total number[a]	Frequency (%)[b]	Interacting pairs[c]	% interacting pairs[d]
K	13446	5.8	5881	43.7
R	10919	4.7	8149	74.6
F	9162	4.0	4969	54.2
Y	8309	3.6	6615	79.6
W	3412	1.5	2446	71.7

Table 3.3: a) The total number of times a particular amino acid appears in the dataset of 593 proteins. b) Occurrence of a particular amino acid in the dataset expressed as a percentage. c) The number of times a particular amino acid occurs as a member of an interacting pair. d) The likelihood of a given amino acid to appear in an interacting pair expressed as a percentage of the number of times that it appears in the dataset.

Amino acid	All cation-π interactions		$E_{es} \leq$ -4.0 kcal/mol	
	Number	Frequency (%)[a]	Number	Frequency (%)[a]
K	1006	7.5	281	2.1
R	1988	18.2	432	4.0
F	915	10.0	199	2.2
Y	1187	14.3	197	2.4
W	892	26.1	317	9.3

Table 3.4: Frequency of cation-π interactions for individual residue types within the 593 protein dataset. a) The likelihood that a particular amino acid type will be in a cation-π interaction, expressed as a percentage of the total number of times that amino acid appears in the dataset. Preferences for arginine and tryptophan are clearly apparent.

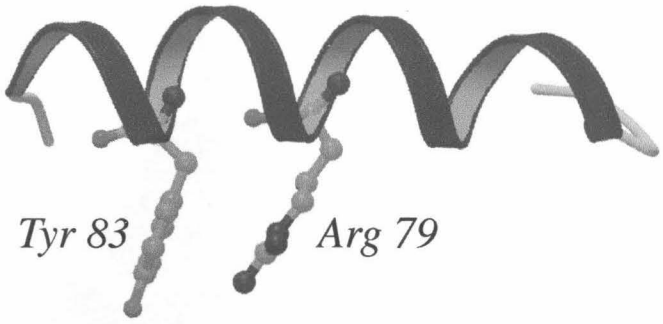


Figure 3.5: An intrahelical cation-π interaction between Arg 79 and Tyr 83 of the vaccinia virus protein VP39 (PDB: 1V39). Plot generated with MOLSCRIPT [35].

interaction than lysine, since arginine is not expected to be innately better at binding to aromatics than lysine. Ab initio calculations by Jorgensen clearly indicate that in the gas phase, ammonium (model for Lys) interacts more strongly with aromatics than guanidinium (model for Arg) [36]. We find at the HF/6-31G** level, the binding energy for ammonium to benzene is -15.3 kcal/mol, substantially stronger than that of guanidinium binding to benzene in both the parallel (-4.1 kcal/mol) and T-shaped (-10.6 kcal/mol) geometries. We expect this pattern to hold up at higher levels of theory. To determine whether our statistics reflect an artifact of our electrostatic models for either Lys or Arg, we modeled each side chain as a point charge, placing a unit charge at either the location of the ϵ-NH_3^+ group of lysine or the location of the guanidinium carbon of arginine. Since these positions essentially overlap in space and the charges are identical, we eliminate any potential bias in the charge model for the cation. We then recalculated E_{es} using this simplified model, and we still find that Arg is substantially more likely than Lys to be found in a significant cation-π interaction. This suggests that other, non-electrostatic effects may be contributing to the propensity of Arg to be involved in cation-π interactions. Since the side chain of arginine is larger and less well water-solvated than that of lysine, it is likely that Arg benefits from better van der Waals interactions with the aromatic ring. In addition, as suggested by Thornton *et al.* [14], the side chain of arginine may still donate several energetically significant hydrogen bonds while simultaneously binding to an aromatic ring (if it is stacked), whereas lysine would typically have to relinquish hydrogen

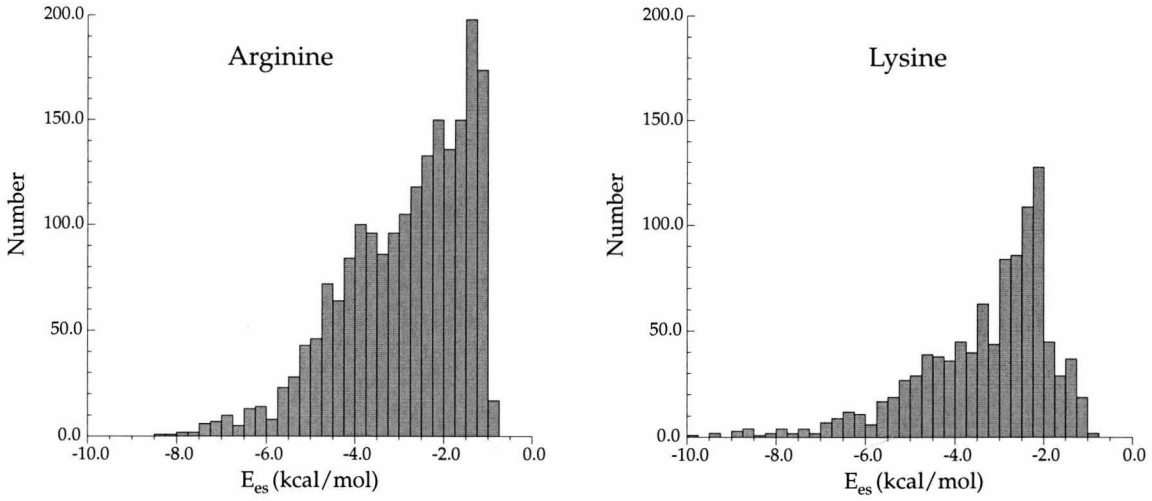


Figure 3.6: Distribution of E_{es} for cation-π interactions involving arginine and lysine.

bonds to water (or other parts of the protein) in order to bind to an aromatic.

Thus, although arginine is more prevalent than lysine in cation-π interactions, we suggest that this does not reflect the intrinsic cation-π binding ability of arginine, but rather other factors as discussed above. Consistent with this view, the average strengths of a cation-π interaction involving either Lys or Arg are similar (Lys: -3.32 $\pm$ 1.50 kcal/mol; Arg: -2.87 $\pm$ 1.42 kcal/mol), with perhaps a slight preference for Lys. This is evident in Figure 3.6 which shows the distribution of E_{es} for cation-π interactions involving arginine and lysine. It is also interesting to note that the 12 most energetically significant interactions involve lysine as the cation, consistent with the ab initio calculations discussed above.

The side chain of arginine may participate in cation-π interactions in two limiting geometries (Figure 3.1). The first is a stacked geometry that places the guanidinium group parallel to the plane of the aromatic ring. The second is a T-shaped geometry that places the guanidinium group perpendicular to the ring. Computational studies have suggested that in the gas phase, the perpendicular geometry is favored energetically, while in solution, the parallel geometry is preferred [36]. Previous studies of arginine-aromatic interactions within proteins by Thornton and co-workers [12, 14], as well as Flocco and Mowbray [18], have suggested that the parallel geometry is preferred in protein structures. Our findings largely agree with these studies, in that

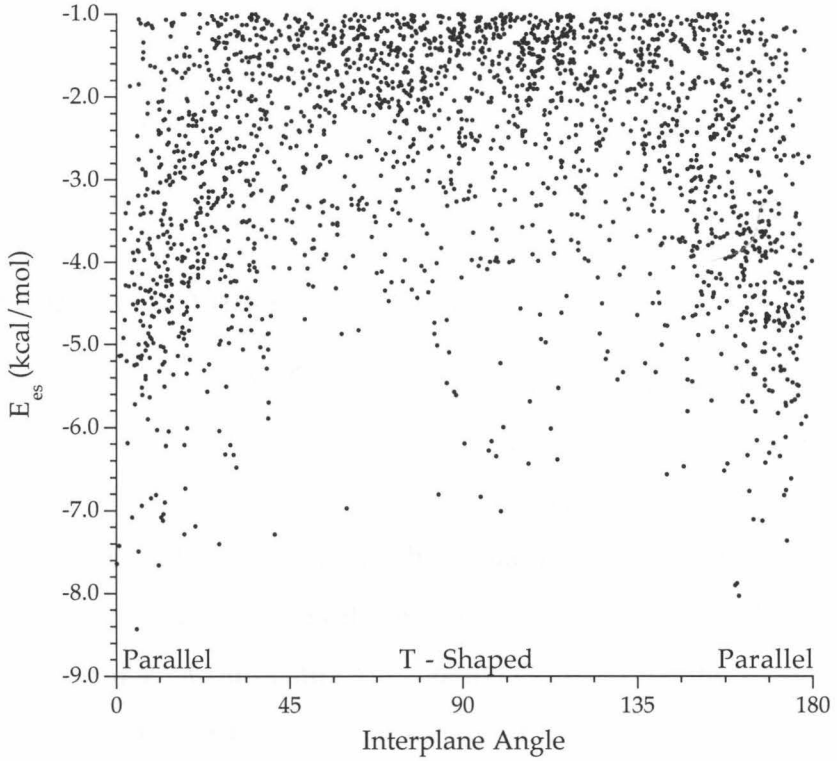


Figure 3.7: The interplane angle of cation-π interactions involving arginine. Although the parallel geometry is preferred, T-shaped geometries represent some of the strongest interactions.

there is a preference for parallel geometries, as shown in Figure 3.7. However, some of the strongest cation-π interactions involve T-shaped geometries. Earlier work proposed that when Arg is in the T-shaped geometry, it always has one, not two, NH_2 groups in the face of the ring. While this is certainly a strong preference, we do find some structures with two NH_2 groups interacting with the aromatic (Figure 3.12).

With respect to the orientation of the lysine side chain in cation-π interactions, the ε-carbon (CE) of lysine is 2.4 times more likely to be closer to the ring centroid than the nitrogen (NZ). This preference, which is contrary to expectations based only on electrostatics, disappears when only the strongest binding structures ($E_{es} \leq$ -5.0

kcal/mol) are considered, although some very strong interactions have a close CE-aromatic contact (Figure 3.11). These results suggest that while the ε-carbon carries substantial partial positive charge, binding to the nitrogen leads to the strongest cation-π interactions. We note, however, that positioning the carbon closest to the ring may contribute favorable van der Waals binding, and exposing the ammonium group may lead to better interactions with solvent or hydrogen bonding groups in the protein.

3.3.2 Aromatic Amino Acids

With respect to the aromatic amino acids, a marked preference for tyrosine and tryptophan to interact with the side chains of lysine or arginine is observed (Table 3.3). On average, 80% of all tyrosines and 72% of all tryptophans in the dataset interact with a cationic side chain, vs. 54% for phenylalanine. To determine whether cation-π interactions explain these observations, the results in Table 3.3 can be compared to those in Table 3.4. Concerning the phenylalanine/tyrosine results, there is no strong bias when considering only the strongest cation-π interactions ($E_{es} \leq$ -4.0 kcal/mol). When all cation-π interactions are considered, however, a slight preference for Tyr is evident. Theory suggests that tyrosine and phenylalanine are equivalent in their cation-π binding ability, and that the increased number of cation-π interactions involving tyrosine must be due to other effects, such as the ability of the OH group of the tyrosine to act as a hydrogen bond donor. It has been shown that if the tyrosine OH is involved in a hydrogen bond, it substantially potentiates the cation-π binding ability of the phenolic ring [32]. Also, the negative electrostatic potential on the oxygen could directly contribute to cation binding, as discussed above. In fact, the OH could accept a hydrogen bond from a cationic side chain, which might position these side chains more favorably so as to benefit from cation-π binding. This is suggested in Figure 3.8, which shows the spatial distribution of cations engaged in cation-π interactions. In the case of Tyr, there appears to be a bias for the cation to be closer to the OH, suggesting that in the proteins, this interaction is energetically

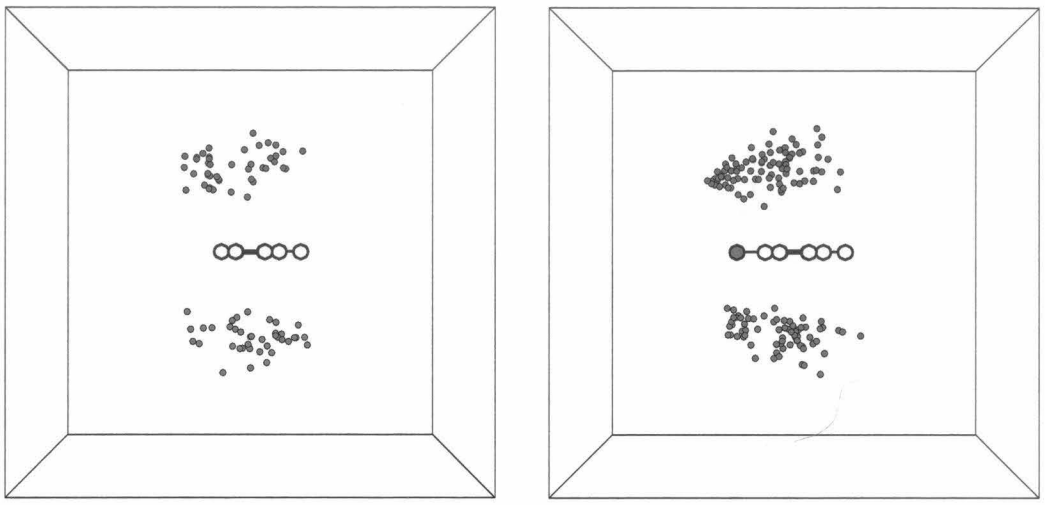


Figure 3.8: Scatter plot for the energetically significant cation-π interactions for Lys/Phe pairs (left) and Lys/Tyr pairs (right) from all 323 single subunit proteins. The small red circles denote the location of the side chain N of the lysines. Each side of the cube is 10 Å. The aromatic rings are viewed edge on, with the β carbons located on the right. The phenol oxygen is in red. Note the cluster of structures above and below the oxygen.

significant.

Perhaps the most surprising result is that 26% of all tryptophans in the dataset are involved in at least one *energetically significant* cation-π interaction. It has been postulated that tryptophan would be overrepresented at cation-π sites, because in the gas phase, indole binds cations more tightly than either benzene or phenol [1, 31, 32]. This preference for indole has a strong electrostatic component, and since our energy calculation mimics this trend, it could be argued that it was inevitable that tryptophan would appear to be a better cation-π binder when using an energetic criterion. To address this concern, two limiting reasons why tryptophan might be more prevalent at cation-π sites were considered. The first is that the larger volume of tryptophan allows it to contact a greater number of cations relative to phenylalanine or tyrosine. To test this, we again consider the purely geometric criterion, counting all cation-aromatic interactions (Table 3.3). We find that cationic side chains show a marked preference to be near tyrosine relative to tryptophan and phenylalanine (Phe, 1.0 : Tyr, 1.47 : Trp, 1.32). When compared to the ratios of Table 3.4, column 2 (Phe, 1.0 : Tyr, 1.43 : Trp, 2.61), it is clear that the Trp bias is not geometric in origin.

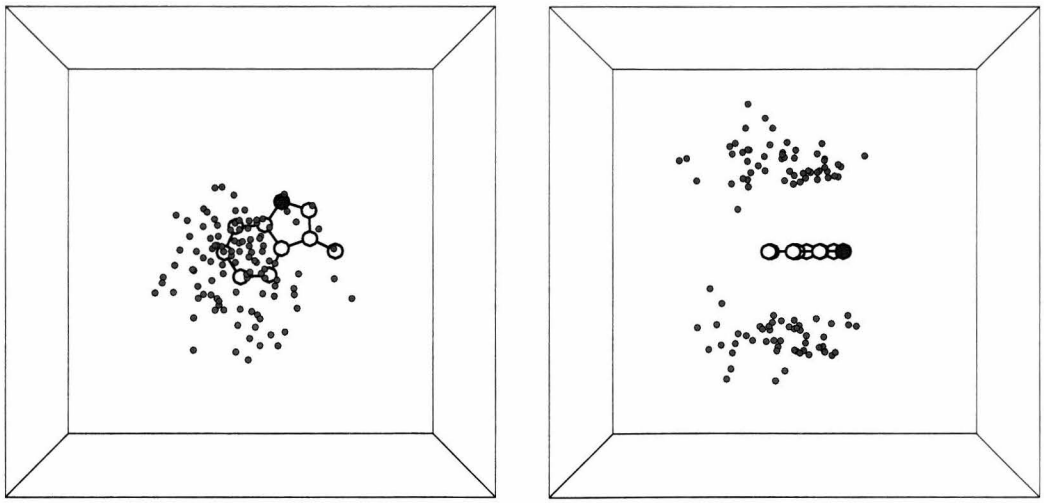


Figure 3.9: Views of all Lys-Trp cation-π interactions from single subunit proteins. The indole nitrogen is shown in blue. Left: Top-down view, right: edge-on view.

To determine whether the bias for Trp simply reflects a bias in our energy model, we reduced tryptophan's cation-π binding ability 50% by halving the electrostatic energy of each interaction involving Trp. This reduction substantially penalizes tryptophan, making it a much less potent cation binder than phenylalanine. Nevertheless, using this different model for Trp, we still find that it is favored over phenylalanine by a ratio of 2.0:1.0. We thus conclude that the substantial overrepresentation of Trp in our collection of cation-π interactions does not result from a bias in our selection criteria. Instead, the likelihood that an aromatic side chain will be involved in a cation-π interaction directly reflects the innate cation-π binding ability of the ring.

Two views of the Lys-Trp interaction are shown in Figure 3.9. A clear bias for the six ring of Trp to be involved in cation-π interactions is evident. This is in complete agreement with expectations based on the electrostatic potential surface of the Trp side chain shown in Chapter 2. Also contributing to this bias is a steric effect; the protein backbone is nearer the five-ring, and may interfere with cation binding. Such an effect is also evident in Figure 3.8, but it is much less pronounced than the bias seen with Trp.

Amino acids	All cation-π interactions		$E_{es} \leq$ -4.0 kcal/mol	
	cation-π	(%) interacting pairs[a]	cation-π	(%) interacting pairs[a]
KF	285	14.5	75	3.8
KY	438	14.7	84	2.8
KW	283	30.2	122	13.0
RF	630	21.0	124	4.1
RY	749	20.6	113	3.1
RW	609	40.4	195	12.9

Table 3.5: Frequency of cation-π interactions for residue pairs within the 593 protein dataset. a) The likelihood that a particular amino acid pair is in a cation-π interaction expressed as a percentage of interacting pairs. Within the dataset, the total numbers of interacting pairs: KF 1962; KY 2981; KW 938; RF 3007; RY 3634; RW 1508.

3.3.3 Pairwise Interactions

Table 3.5 considers the interactions between the six possible cation-π pairs. When the side chain of lysine approaches that of either phenylalanine or tyrosine, there is an approximately 15% chance they will be involved in an energetically significant cation-π interaction. However, when lysine approaches tryptophan, the chances are nearly one in three that the residues will be involved in a cation-π interaction. These trends also hold for interactions involving arginine. Again the preference for tryptophan stands out.

An interesting question concerns the location of cation-π interactions within protein structures. It is generally accepted that cationic residues prefer to be on the surface of a protein, whereas aromatic amino acids prefer to remain in the hydrophobic core. Since a cation-π interaction within a protein contains both a cation and an aromatic, it is worth considering whether the interacting pairs prefer to be located on the surfaces of proteins, or in the cores. Traditional methods for determining residue surface accessibility rely on calculating the surface area of an amino acid that is in contact with water. Since the partners of a cation-π interaction are necessarily in contact with one another, their water accessible surface is diminished, even though the interacting pair *as a unit* may be fully solvated. Thus, it is difficult to determine computationally whether a cation-π interaction is on the surface of a protein using only surface accessibility. We have not visually inspected all 2,994 cation-π interac-

tions, but examination of many structures suggests that cation-π interactions tend to be on the surfaces of proteins, consistent with an earlier conclusion by Flocco and Mowbray [18]. We will revisit this subject in further detail in Chapter 4.

Although the cationic side chains of arginine and lysine often experience favorable electrostatic interactions with the side chains of aromatic amino acids via the cation-π interaction, the question remains whether Nature uses this advantage to orient these side chains in folded proteins. That is, when all interactions between a cationic and an aromatic side chain are considered, is there a tendency for geometries that are favorable in a cation-π sense compared to just a random distribution? To answer this question, we consider the simplest system – lysine /phenylalanine – and ask whether the side chains of these residues are placed such that the lysine is in an electrostatically favorable location above the phenylalanine ring. To do this, we must define what a random distribution would look like. While others have based such a distinction on criteria such as interplane angle, this approach was somewhat ambiguous for our purposes. So, we sought a clear geometric approach. Our approach is best understood with reference to Figure 3.10.

We first create a model for benzene that describes its size and shape, as well as the volume excluded by the benzene and atoms touching it. We also create a "shell" with the same shape of the benzene, but a uniformly larger size, and we consider only structures for which the Lys NH_3^+ lies within this shell. We then create a cylinder encompassing the excluded volume of benzene and the region above it. If a cation lies within this cylinder, it is considered to be in an electrostatically favorable region and so is expected to experience a cation-π interaction. The cylindrical region above the ring occupies 32% of the total volume, suggesting that 32% of the Lys should be in this region. However, we find that 48% of the lysines lie in this cylinder. With a dataset of 1,716 structures,[4] this indicates a non-random distribution at a confidence level greater than 99.999%. Similar trends are observed for the other cation-π pairs. Thus, proteins do position cations at non-random positions relative

[4]This number differs from the value in Table 3.5 (1,962) because many legitimate cation-π interactions have the N outside of the shell, but CE is inside the shell.

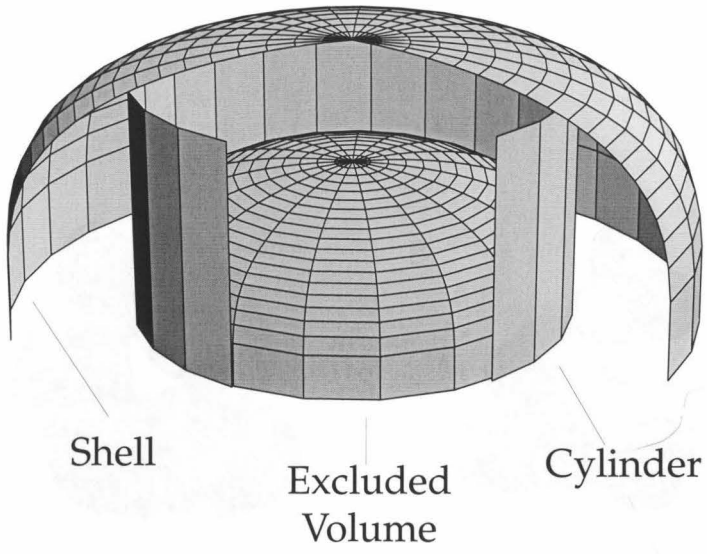


Figure 3.10: Schematic view of model used to calculate the preferred location for lysine/phenylalanine pairs. The excluded volume region represents the volume of benzene plus the volume unavailable to atoms of radius 1.7 Å. The cylinder is tangent to the benzene (radius = 4.8 Å obtained by adding the radius of benzene (1.4 Å + 1.7 Å) and the radius of a neighboring carbon atom (1.7 Å)). The shell is obtained by adding a constant radius of 2.8 Å – the diameter of a water molecule – to the excluded volume region. This view shows the only the top half of the excluded volume and has portions of the shell and cylinder removed for clarity. To determine whether there is a preference for cations to be located within the cylindrical region, we compare the number of lysine/phenylalanine structures located within the cylindrical region, to the total number of lysine/phenylalanine structures inside of the shell.

to aromatics, so as to optimize cation-π interactions, providing further support for the notion that cation-π interactions contribute significantly to protein stability. Although the methodology differs, our results agree with the findings of Thornton *et al.* [17] who observe non-random angle preferences for cationic side chains interacting with aromatic side chains.

3.4 Gallery of Cation-π Interactions

Finally, we present a gallery of representative cation-π interactions obtained in this study, along with their E_{es} values.

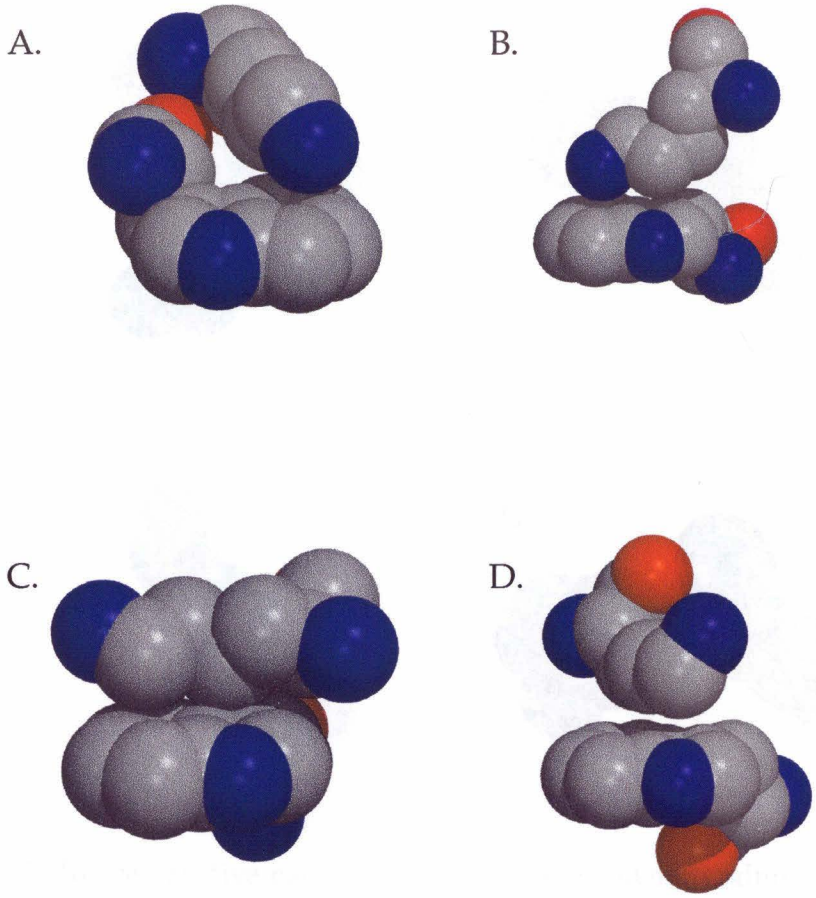


Figure 3.11: Representative cation-π interactions involving lysine. *A.* Interaction between Lys379A and Trp376A in a copper amine oxidase (PDB: 1SPU). E_{es} for this interaction is -9.8 kcal/mol, making it the strongest in the dataset. *B.* Interaction between Lys408 and Trp356 in a sulfite reductase (PDB: 1AOP). E_{es} for this interaction is -9.4 kcal/mol, making it the third strongest in the dataset. *C.* Interaction between Lys133 and Trp140 in Staph. nuclease (PDB: 2SNS). E_{es} for this interaction is -7.3 kcal/mol. It is the strongest interaction in the dataset that places the methylene of the lysine closer to the aromatic than the ammonium group. *D.* Interaction between Lys76B and Trp58B in transaldolase B. (PDB: 1ONR). E_{es} for this interaction is -4.4 kcal/mol. It also places the methylene of the lysine closer to the aromatic than the ammonium group.

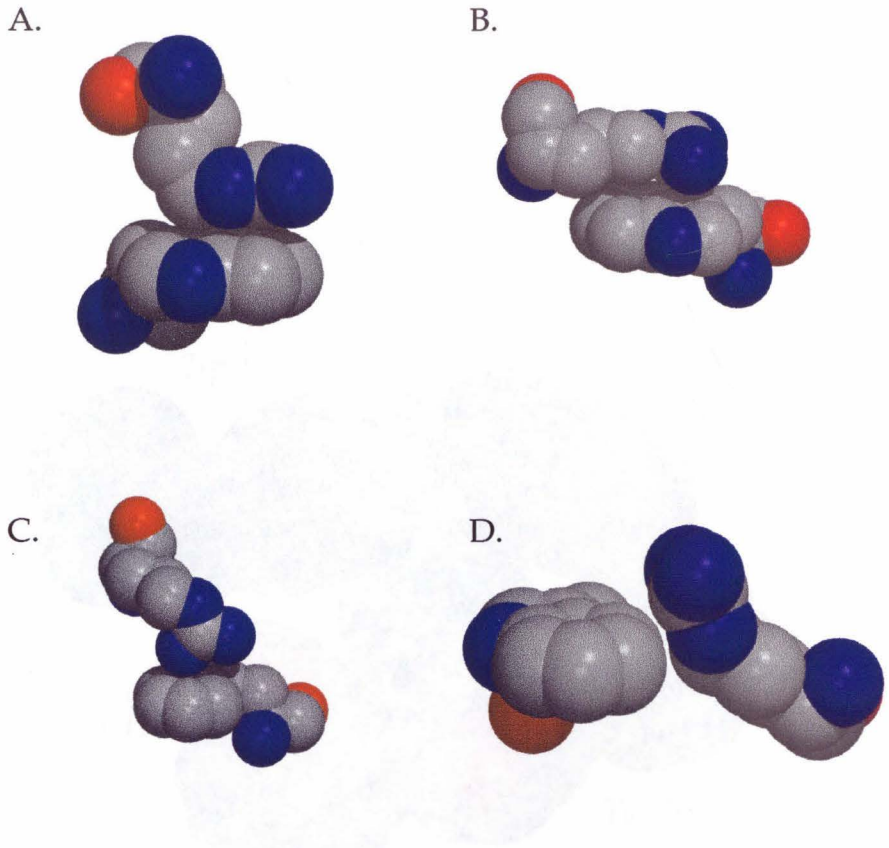


Figure 3.12: Representative cation-π interactions involving arginine. *A.* Interaction between Arg77A and Trp211A in the oligopeptide binding protein OPPA (PDB: 1JET). E_{es} for this interaction is -8.4 kcal/mol, making it the strongest interaction involving arginine in the dataset. When E_{vdW} is included, this is the strongest interaction overall. *B.* Interaction between Arg335 and Trp481 in a mutant of glucose-6-phosphate dehydrogenase (PDB: 2DPG). E_{es} for this interaction is -7.7 kcal/mol. It is an example of a strong stacked arginine interaction. *C.* Interaction between Arg1136 and Trp1175 in the tyrosine kinase domain of the insulin receptor (PDB: 1IRK). E_{es} for this interaction is -6.8 kcal/mol. Two nitrogens face the aromatic ring in this T-shaped structure. Thornton has noted that this is an uncommon motif [14]. *D.* Interaction between Arg1253 and Trp1193 in the tyrosine kinase domain of the insulin receptor (PDB: 1IRK). E_{es} for this interaction is -1.0 kcal/mol. It is the weakest interaction that we accept. E_{vdW} for this interaction is -2.0 kcal/mol.

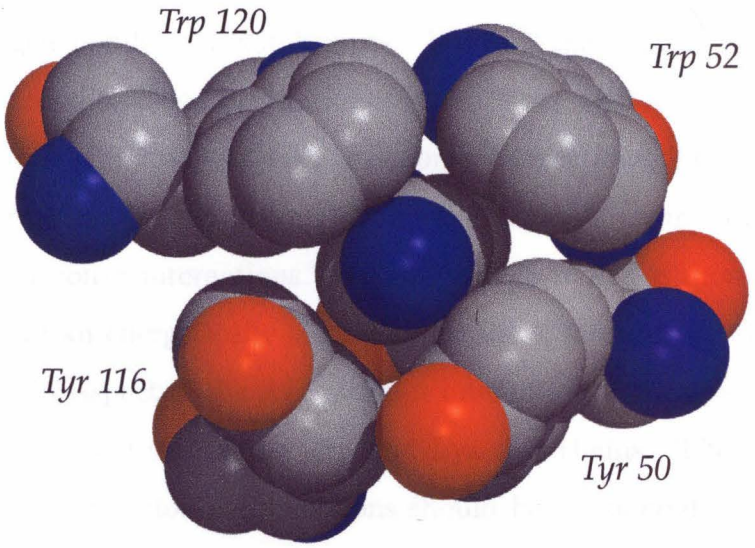


Figure 3.13: A remarkable example of a lysine involved in 4 cation-π interactions taken from the protein glucoamylase (PDB: 1GAI). The total E_{es} for this interaction is $\approx$ -22 kcal/mol.

3.5 Summary

By developing an energy-based criterion that puts all cation-π interactions on an equal footing, we have been able to develop meaningful statistics for the frequency of occurrence of cation-π interactions in proteins and to evaluate whether specific cation-π pairs are preferred. We find that cation-π interactions are common – one favorable interaction can be expected for every 77 residues of protein length. While the weakest of the interactions considered here ($E_{es} \approx$ -1.0 kcal/mol) may make only a small contribution to the overall stability, it is clear that some of the more favorable pairs contribute at least as much to protein stability as the more conventional interactions such as hydrogen bonds and salt bridges. We find that Trp is the most likely of the aromatics to be involved in a cation-π interaction, with a remarkable 26% of all Trps involved in energetically significant cation-π interactions. This is consistent with theoretical arguments [1, 32] that predicted the Trp side chain would be especially well suited to cation-π interactions. We also find that Arg is more likely than Lys to be involved in an energetically significant cation-π interaction. This is likely not due to an intrinsic aspect of the Arg cation-π interaction, but more likely reflects the differing geometric features of the Arg and Lys side chains. These results make a compelling case that cation-π interactions should be considered alongside the more conventional hydrogen bonds, salt bridges and hydrophobic effects in any analysis of protein structure.

3.6 Appendix 1 – CaPTURE Manual

3.6.1 Overview

Cation-π interactions are widely distributed throughout protein structures. CaPTURE (Cation-π Trends Using Realistic Electrostatics) is a program that allows the user to search Protein Data Bank (PDB) files to find *energetically significant* cation-π interactions. Using user specified parameters, CaPTURE searches a PDB file for potential cation-π interactions using geometric criteria, and evaluates them energet-

ically using a modified version of the OPLS force field. The user can specify both the geometric and energetic criteria that define a cation-π interaction. CaPTURE supports both an interactive mode which leads the user through the process, and a batch mode that lends itself to quick, large-scale database searching. For a study presenting results obtained by CaPTURE, see: "Cation-π Interactions in Structural Biology," J.P. Gallivan and D.A. Dougherty, *Proceedings of the National Academy of Sciences, USA*, **1999**, *96*, 9459.

3.6.2 Installation

CaPTURE is distributed as C++ source code in the file `capture.tar`. CaPTURE has been compiled on SGI workstations running IRIX 6.2 and IRIX 6.5.5. In addition, CaPTURE has been run with minimal modifications under both Solaris and Linux. Although there is no reason to suspect that it won't compile under other systems, the author does not the have the time or resources to explore this possibility. If you succeed in compiling CaPTURE on another system, congratulations! I would love to hear of your experiences, and with your permission, I can make your distribution available.

To install on an SGI system, simply untar the distribution file and type make:

```
tar -xvf capture.tar
make
```

There will likely be several warnings, but hopefully no fatal errors. The finished product is an executable named `capture` which is about 450K. CaPTURE also uses a configuration file that will be generated the first time the program is run, or anytime that the program can't find it.

3.6.3 Getting Started

CaPTURE supports two modes of operation, an interactive mode and a batch mode. The first time the program is run, it will enter the interactive mode and prompt the user to answer several questions. When the questions are answered, a configuration file (`capture.cfg`) is generated that can be used in the future. Once this file is present and is to the user's liking, the program can be run in batch mode, bypassing the Q&A.

Running CaPTURE in Interactive Mode

To run CaPTURE in the interactive mode, simply type '`capture`'. The program will prompt you for a PDB filename. Once the filename is entered, CaPTURE will ask the user whether it should use default parameters or custom parameters to carry out its search. In most cases, the default parameters are acceptable. Descriptions of the parameters appear later in this document. The default parameters were used in our published study: "Cation-π Interactions in Structural Biology," J.P. Gallivan and D.A. Dougherty, *Proceedings of the National Academy of Sciences, USA*, **1999**, *96*, 9459.

Running CaPTURE in Batch Mode

To run CaPTURE in batch mode, simply type '`capture` *pdbfilename*'. Provided both the pdb file and the configuration file (`capture.cfg`) are present, the program will run without the need for further user intervention.

3.6.4 Interpreting the Screen Output from CaPTURE

When the program starts, you should see the following:

```
Reading the PDB file filename
Reading parameters from the file capture.cfg.
Initializing.
```

At this point, there may be some warnings printed:

```
********** Amino Acid Warnings **********
WARNING! Atoms in the following amino acids have fractional occupancy <
0.95.
```

This means that there is uncertainty in their positions. These amino acids will be excluded from the calculations.

```
.
.
Atom 68 in residue THR 9 has an occupancy < 0.95 and is excluded.
Atom 69 in residue THR 9 has an occupancy < 0.95 and is excluded.
Atom 70 in residue THR 9 has an occupancy < 0.95 and is excluded.
Atom 71 in residue THR 9 has an occupancy < 0.95 and is excluded.
Atom 72 in residue THR 9 has an occupancy < 0.95 and is excluded.
Atom 73 in residue THR 9 has an occupancy < 0.95 and is excluded.
Atom 304 in residue LYS 38 has an occupancy < 0.95 and is excluded.
Atom 305 in residue LYS 38 has an occupancy < 0.95 and is excluded.
Atom 306 in residue LYS 38 has an occupancy < 0.95 and is excluded.
Atom 307 in residue LYS 38 has an occupancy < 0.95 and is excluded.
Atom 308 in residue LYS 38 has an occupancy < 0.95 and is excluded.
.
.
********** End of Amino Acid Warnings **********
```

In particular, if atoms in the PDB file have fractional occupancies less than 0.95, they will be excluded from the calculations. This is because there is some uncertainty in their positions. This is usually not a serious problem, unless there are a very large number of warnings. In addition, many of these atoms may belong to residues that cannot be involved in a cation-π interaction.

You will then see something like the following:

```
********** Report for the protein filename **************
This protein has 108 amino acids of which:
3 (2.8%) are Lysine
3 (2.8%) are Arginine
4 (3.7%) are Phenylalanine
6 (5.6%) are Tyrosine
3 (2.8%) are Tryptophan
This protein exists as a single chain
The protein has an apparent molecular weight of 12151 D
```

These are largely self-explanatory. You may also see something like:

```
WARNING! These amino acids are missing atoms needed for the calculations:
LYS 38
```

The warning in this case means that the program knows that the amino acid is there, but some of the atoms may have a fractional occupancy of < 0.95, or may be missing from the PDB file altogether. Amino acids listed here are only those potentially involved in cation-π interactions or salt bridges. Note that there may be amino acids in this list that are not listed in the amino acid warnings above (for example, a lysine that does not have its ϵ-nitrogen in the PDB file may appear in this list, but not above).

The following section summarizes the major findings.

```
********************** Cation-Pi Summary ***********************
Number of ARG/PHE interacting pairs:  1
Number of energetically significant ARG/PHE cation-pi interactions:  1
The following are the energetically significant ARG/PHE cation-pi
interactions:
```

```
 Cation AA   #   Chain  Pi AA   #   Chain     E(es)       E(vdw)
                                            (kcal/mol)  (kcal/mol)
    ARG     92          PHE    21            -1.84       -1.07

Number of ARG/TYR interacting pairs:  0
Number of energetically significant ARG/TYR cation-pi interactions:  0
Number of ARG/TRP interacting pairs:  0
Number of energetically significant ARG/TRP cation-pi interactions:  0
Number of LYS/PHE interacting pairs:  0
Number of energetically significant LYS/PHE cation-pi interactions:  0
Number of LYS/TYR interacting pairs:  1
Number of energetically significant LYS/TYR cation-pi interactions:  0
Number of LYS/TRP interacting pairs:  0
Number of energetically significant LYS/TRP cation-pi interactions:  0
* There is a total of 1 energetically significant cation-pi
interactions.
* There are 0 cation-pi interactions with E(es) <= -2.0 kcal/mol
* There are 1 cation-pi interactions with E(es) between -2.0 and -1.0
kcal/mol and with E(vdw) <= -1.0 kcal/mol
* There is an average of 1 cation-pi interaction for every 108.0
residues.
Successful Exit
```

An interacting pair is a pair of amino acids in which the appropriate atoms of the side chains are arranged such that a water molecule cannot fit in between at the point of closest approach.

Using the default parameters, an energetically significant cation-π interaction is an interacting pair whose OPLS electrostatic energy, E_{es}, is either $\leq$ -2.0 kcal/mol OR whose E_{es} is $\leq$ -1.0 kcal/mol AND whose OPLS van der Waals energy, E_{vdW}, is $\leq$ -1.0 kcal/mol. The criteria used for these calculations may be changed by the user.

For examining a small number of proteins, the screen output provided by the default parameters is normally quite sufficient for discovering cation-π interactions.

However, for analysis of large numbers of proteins, or for more details regarding the interactions, you may want to try to adjust the some of the parameters for tailoring your search.

3.6.5 Adjusting the Parameters

The simplest way to change the parameters is to run the program from the command line without specifying a pdb filename. Simply type '`capture`'. You will then be given the option to change the parameters. This section will describe each of the options presented.

The first questions involve energy cutoffs for determining cation-π interactions.

```
We use a 3-tiered system to determine whether an interaction is
significant.
Tier 1:  Interactions with E(es) below a primary cutoff (-2.0 kcal/mol)
are automatically counted.
Tier 2:  Interactions with E(es) below a secondary cutoff (-1.0 kcal/mol)
are counted only if E(vdW) is less than a given energy
(-1.0 kcal/mol).
Tier 3:  Interactions with E(es) above the secondary cutoff are rejected.
What OPLS electrostatic energy, E(es), in kcal/mol should be used
for the primary cutoff (Tier 1)?
(Default = -2.0)
->
What OPLS electrostatic energy, E(es), in kcal/mol should be used
for the secondary cutoff (Tier 2)?
(Default = -1.0)
->
What OPLS van der Waals energy, E(vdW), in kcal/mol should be used
in conjunction with the secondary cutoff (Tier 2)?
(Default = -1.0)
```

```
->
```

The next question asks whether CaPTURE should consider salt bridges. In this context, a very liberal definition of a salt bridge is used. An interaction is considered a salt bridge when a cationic residue (Arg or Lys) forms an interacting pair with an anionic residue (Asp or Glu). As described above, an interacting pair is a pair of amino acids in which the appropriate atoms of the side chains are arranged such that a water molecule cannot fit in between at the point of closest approach.

```
Would you like to consider salt-bridges also?
With this option, you will only see interactions determined by distance.
Energy calculations are not performed.
(Enter 1 for yes or 0 for no.  Default = 0)
->
```

The next question asks whether CaPTURE should write a PDB file showing only the two amino acids in a given interaction. Be forwarned that there may be many files written, which may take up a lot of space.

```
Would you like a PDB file describing each interaction to be written?
(Enter 1 for yes or 0 for no.  Default = 0)
->
```

The next question asks whether you would like to see the advanced options. For most purposes, these need not be changed. However, the option to write '.stat' files (described below) found within the advanced options may be useful to some users.

```
Would you like to see advanced options?
(Enter 1 for yes or 0 for no.  Default = 0)
->
```

The following questions all deal with the geometrical definitions used in the initial screen for cation-π interactions. Since CaPTURE uses energy-based criteria for determining cation-π interactions, these would typically not be changed. Increasing the distance beyond 10Å, for example, would likely not result in finding more energetically significant cation-π interactions, but would increase the calculation time by forcing the program to calculate the energies of interactions that are unlikely to be significant. In addition, the angle preferences are of little use, unless you would like to use the program for more specific searches.

```
What is the minimum distance (in angstroms) between interacting pairs?
(Default = 0.0)
->
What is the maximum distance (in angstroms) between interacting pairs?
(Default = 10.0)
->
What is the minimum theta angle (in degrees) between interacting pairs?
(Default = 0.0)
->
What is the maximum theta angle (in degrees) between interacting pairs?
(Default = 360.0)
->
What is the minimum phi angle (in degrees) between interacting pairs?
(Default = 0.0)
->
What is the maximum phi angle (in degrees) between interacting pairs?
(Default = 360.0)
```

```
->
```

The next question asks whether CaPTURE should write a file showing the OPLS energies for each interaction. Each file will contain the pairwise interaction energies for all atoms considered. Be forwarned that there may be many files written, which may take up a lot of space. This option is not very useful, except for debugging.

```
Should I write an output file detailing the OPLS energies for each
interaction?  These can take up a lot of space.
(Enter 1 for yes or 0 for no.  Default = 0)
->
```

The final question asks whether CaPTURE should write an output file showing the statistics calculated by the program in a tabular form. The resulting '.stat' file (called pdbfilename.stat) can be processed easily using programs such as *awk* or *perl* to extract data, or by using a spreadsheet program to read the file. A description of the '.stat' files appears later in the document.

```
Should I write an output file containing all of the statistics
in tabular form?  This may be hard to parse, but is useful for large
scale searching.  (Enter 1 for yes or 0 for no.  Default = 0)
->
```

3.6.6 Limitations of CaPTURE

As distributed, CaPTURE has a few limitations for the sizes of proteins that it can use. In normal use, these are unlikely to be a problem. CaPTURE can handle proteins

of up to 100,000 atoms, with up to 2,000 residues per chain, and up to 30 chains. Currently the maximum number of interactions is set at 5,000. All of these values can be changed by editing the file '`global.h`' and recompiling.

CaPTURE cannot currently deal with non-protein units such as ligands, post-translational modifications, and nucleic acids. All of these are ignored by the program. Because of the difficulty in determining the appropriate OPLS parameters for heteroatoms, there are no plans to introduce support for these.

CaPTURE's PDB parser has recently been improved, but is certainly not perfect. If you have a PDB file that causes the program to crash, you may try to 'convert' it using the freely available program *Babel*. Simply convert the file from a PDB to a PDB. This will rewrite the file in a standard form that may be easier to read in.

CaPTURE cannot handle inserted residues within PDB files. An example occurs in the protein crambin (PDB: 1CBN), where residue 22 is a proline and residue 22B is a serine. CaPTURE will warn whether a protein has inserted residues and gracefully exit.

Bugs

Although CaPTURE has been extensively tested, there are bound to be some bugs in the program. If you spot one, please let me know by email (jpg@igor.caltech.edu) and I will take a look at it. CaPTURE has been tested on hundreds of files from the PDB and works reasonably well. As with any computational package, you should make sure that the results from the program pass the common sense test. If the program produces seemingly non-sensical results, please let me know and I will try to track down the problem.

Notes

The configuration file, `capture.cfg`, is created by the program the first time it is run. It contains all of the information that CaPTURE needs to do its job other than the name of the PDB file. You may edit this file, but it must NOT contain any blank lines or the program will hang. Additionally, there are options in the file that are not supported in the public distribution of CaPTURE (such as calculating solvent accessible surface areas). It is best that you not modify this file by hand. If you want

to modify it, you can run CaPTURE in the interactive mode and choose to change the parameters. This will update the configuration file.

Contents of the '.stat' file

For large-scale searching, it is often useful to have the output of the program written to a file in tabular form. This can be set under the advanced options setting (see above). When this option is set, CaPTURE will produce a '.stat' file, with the name '`pdbfilename.stat`' where pdbfilename is that of the pdb file entered. The '.stat' file has four main sections: a remark section (REM); a sequence section (SEQ); a warning section (WARN); and an interaction section (INT). Descriptions of these sections appear below.

Remark Section (REM):

This consists of a single line in the file (the first line) that begins with 'REM:'. This line contains the name of the original PDB file.

Sequence Section (SEQ):

This section is a single line beginning with the word SEQ. This line is tab-delimited into 'fields'. The first field contains the word SEQ. The second contains the filename. The following fields contain the number of residues in the protein, followed by the number of times each amino acid (limited to K,R,F,Y,W,D, and E) occurs in the protein, as well as the percent composition of each amino acid.

Warning Section (WARN):

This section is a single line beginning with the word WARN, followed by the file name. It reports whether amino acids are missing atoms needed for the calculations. For example, sometimes only the backbone trace of a protein will be in the PDB file. Even though the protein may have 3 lysines in its sequence, there are no coordinates for the side chains and thus the program cannot determine whether they are in cation-π interactions. These amino acids also appear under the '`WARNING! These amino acids are missing atoms needed for the calculations:`' section in the screen output. Also on this line are the percentages of the amino acids that are missing atoms. For example, if a protein has 3 lysines in the sequence, and one is missing atoms, this number will be $1/3 = 33.3\%$. If the percentages are very high, this indi-

cates a PDB file that is missing a significant number of atoms, and care should be taken in interpreting the output of the program.

Interaction Section (INT):

This section is the most important of the '.stat' file. It contains information about each pairwise interaction considered. Each line consists of tab-delimited fields that are as follows:

1. The word INT.
2. A representation of the filename of the interaction. If you have specified the option of writing PDB files, this will be the filename of the PDB file for the interaction (if you replace the .z with a .pdb).
3. The type of cation.
4. The residue number of the cation.
5. The chain identifier of the cation (or a dash if the protein is a single chain).
6. The type of π or anion.
7. The residue number of the π or anion.
8. The chain identifier of the π or anion (or a dash if the protein is a single chain).
9. The distance, r, in Å between the cationic center and either the centroid of the aromatic, or the carbonyl carbon of an anionic side chain.
10. The angle, θ, in degrees, between a vector from the ring centroid and the cationic center, and a vector between the β-carbon of the π-system and the centroid. This is undefined for an anionic partner and will appear as -1000.0.
11. The dihedral angle, ϕ, in degrees, between a vector from the ring centroid and the cationic center, and a vector between the β-carbon of the π-system and a normal to the ring. This is undefined for an anionic partner and will appear as -1000.0.

12. The OPLS van der Waals energy, E_{vdW}, (in kcal/mol), for the interaction. Note that this is not applicable for salt bridges.

13. The OPLS electrostatic energy, E_{es}, (in kcal/mol), for the interaction. Note that this is not applicable for salt bridges.

14. The total OPLS energy (in kcal/mol), for the interaction. This is the sum of the van der Waals and electrostatic energies (12 + 13). Note that this is not applicable for salt bridges.

15. The interplane angle between an arginine side chain and a ring. 90 degrees is perpendicular (T-shaped), and 0 and 180 are parallel.

16. The 'gap' measurement. If this number is less than 0, no water molecule can fit between the partners at their closest approach point.

17. This is r as specified in 9, but it is calculated relative to the 6-ring in the case of Trp. This makes plotting easier.

18. This is θ as specified in 10, but it is calculated relative to the 6-ring in the case of Trp. This makes plotting easier.

19. This is ϕ as specified in 11, but it is calculated relative to the 6-ring in the case of Trp. This makes plotting easier.

20. This is the distance between either the ε-carbon of Lys or the δ-carbon of Arg and the centroid of the closest aromatic ring. This will appear as -1000.00 for a salt bridge.

21. The total number of residues in the protein.

Using standard UNIX tools such as *grep*, *cut*, and *awk*, you can easily write scripts to extract information from these files. Alternatively, they may be read into spreadsheets.

3.7 Appendix 2 – Description of CaPTURE Source

This section is meant to describe the CaPTURE source code so that others may be able to modify it. The various features of many of the functions will be briefly outlined.

CaPTURE is written in C++ and was developed under IRIX 6.2 and IRIX 6.5.5. There are no graphics or IRIX-specific functions within the code. Thus, CaPTURE should be able to be compiled under other systems with minimal modifications.

The main body of code is found in the source file `Capture.c`. A schematic of the files and functions called by `Capture.c` are shown in Figure 3.14. Descriptions of the major functions in each source file follow.

Capture.c

`Capture.c` is the heart of the program. Many of the structures and variables are initialized within `Capture.c`.

Welcome.c

`Welcome.c` introduces the user to the program. It first checks whether the user entered a PDB filename on the command line. If so, it checks for the existence of the configuration file `capture.cfg`. If `capture.cfg` exists, the program uses the parameters stored within the configuration file and runs without further user input. If either a PDB filename or the configuration file are missing, the program enters interactive mode, and prompts the user for more information. If there is no configuration file, the WriteConfig function (defined in `WriteConfig.c`) is called, and creates the file `capture.cfg`.

ReadPDB.c

`ReadPDB.c` reads information from the PDB file into the `PDB` structure. It also determines the total number of chains within the protein.

ReadParams.c

`ReadParams.c` assigns default values to many of the user-changeable parameters. It then reads information from the configuration file `capture.cfg` and adjusts the parameters accordingly. There are a few parameters in here, such as the options for

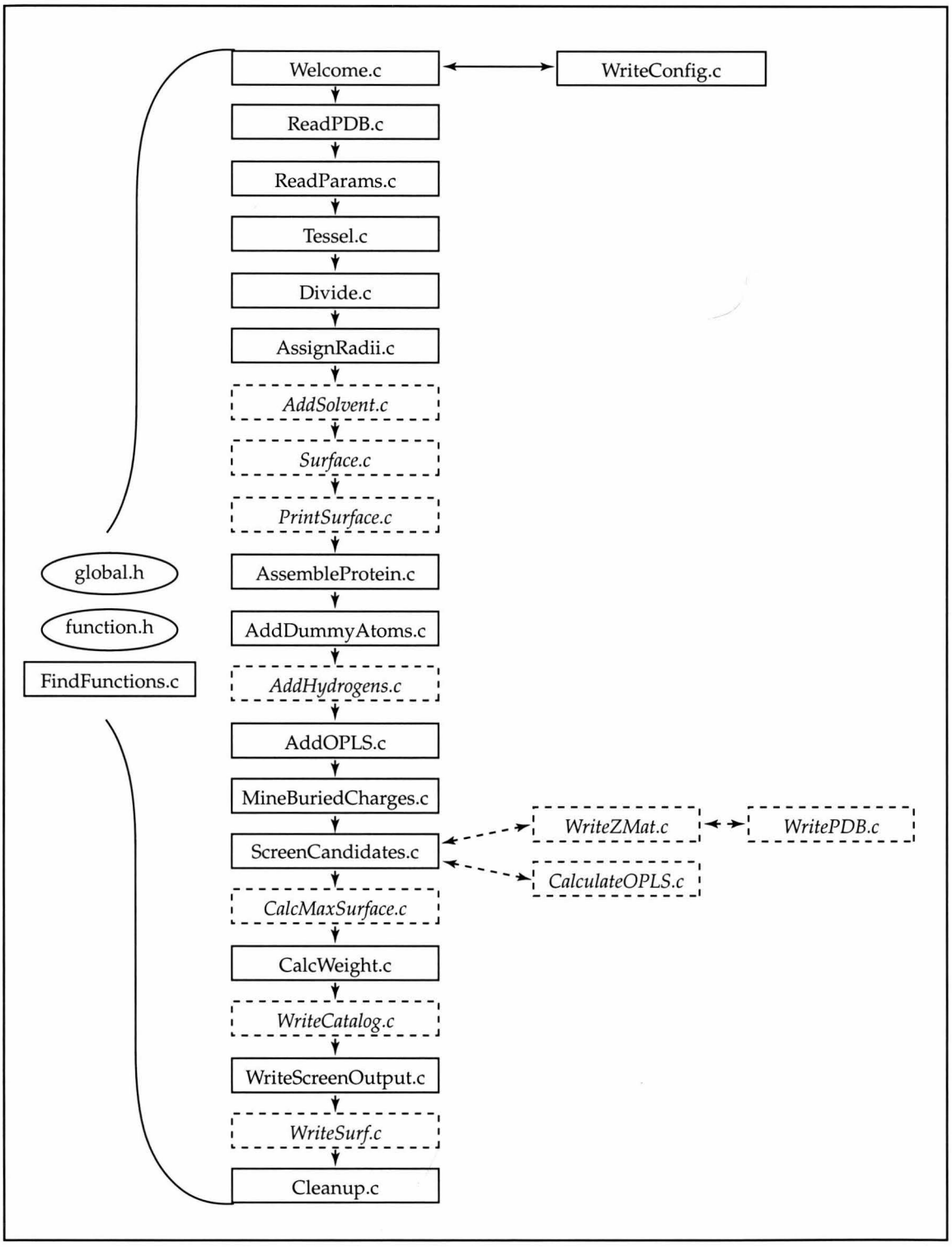


Figure 3.14: Flowchart of `Capture.c` The header files (in ovals) are available to all functions, as are the functions found in `FindFunctions.c`. Files shown in solid boxes are always called. Those in dashed boxes are optional. The arrows show the direction of program flow.

Gaussian calculations, which can only be altered at compile time.

Tessel.c and Divide.c

`Tessel.c and Divide.c` contain functions used by the GEPOL algorithm for calculation of the solvent-accessible surface area of the protein. (See: Silla, E.; Tuñón, I.; Pascual–Ahuir, J. L. *J. Comput. Chem.* **1991,** *12,* 1077–1088.) They are small functions and are called regardless of whether the solvent-accessible surface is actually calculated. These should not be modified.

AssignRadii.c and AddSolvent.c

`AssignRadii.c` assigns an atomic radius to each atom for use in calculating the solvent-accessible surface area. It uses the values from: Rose, G. D.; Geselowitz, A. R.; Lesser, G. J.; Lee, R. H.; Zehfus, M. H. *Science* **1985,** *229,* 834–838. If the solvent-accessible surface is to be calculated, the `AddSolvent` function adds the user-defined solvent radius to each atom.

Surface.c and PrintSurface.c

These files contain the code to determine the solvent-accessible surface of each atom. They are adapted from: Silla, E.; Tuñón, I.; Pascual–Ahuir, J. L. *J. Comput. Chem.* **1991,** *12,* 1077–1088.

AssembleProtein.c

Until now, all data are treated as individual atoms. This function parses the atoms and creates the `Protein` structure. A protein consists of chains, which consist of amino acids, which consist of a series of pointers to the atoms in the `Atom` structure. This function also determines whether the occupancy of the individual atoms is greater than a user-specified value. In addition, the relative accessibility of each amino acid can be calculated.

AddDummyAtoms.c

This function adds dummy atoms to each aromatic ring, including the centroids and surface normals used to calculate various geometric parameters.

AddHydrogens.c

This function adds hydrogen atoms to each aromatic ring (provided they are not there already) for use in calculating the OPLS energies. Since Arg and Lys are treated

as united atoms, hydrogens are not added.

AddOPLS.c

This function adds the OPLS parameters, q, σ, and ϵ to the atoms using the `AssignOPLS` function, defined within `AddOPLS.c`. The parameters are taken from [23, 24].

MineBuriedCharges.c

This function is largely a relic and should probably be phased out. It was originally intended to determine whether an amino acid is buried by comparing its solvent-accessible surface area to a user-defined parameter. It also calculates the total charge of a protein and the population of each type of amino acid.

ScreenCandidates.c

The `ScreenCandidates` function is the computational engine of the CaPTURE program. `ScreenCandidates` searches through the protein for potential cation-π interactions, or salt bridges. When it finds a potential pair of residues, it checks to see if all of the necessary coordinates are present, and uses the functions defined in `FindFunctions.c` to calculate the distance, angle, and dihedral information. Using this information, it can call the functions `Write*ZMatrix.c` and `WritePDB.c` to write files describing the interactions. `ScreenCandidates` can also call the function `CalculateOPLS` (defined in `CalculateOPLS.c`), to determine the OPLS energy of the interacting pair. If the interacting pair is between an Arg and an aromatic, the interplane angle (IPA) is calculated. In addition, `ScreenCandidates` performs various book-keeping activities, including counting the total number of each interaction, and creating the filenames for the .Z and .PDB files.

CalcMaxSurface.c

If CaPTURE is set to calculate the solvent-accessible surface area, this function calculates the maximum accessible area of a given residue (i) by calculating its area in the presence of the backbone coordinates of the $(i-1)$ and $(i+1)$ residues.

CalcWeight.c

This function calculates the molecular weight of the protein.

WriteCatalog.c

This function writes the '`.stat`' file which contains all of the data generated by the program in tabular form, useful for large-scale database searching.

WriteScreenOutput.c

This function formats the most important results of the program for display on the screen. In many cases, this is the only output a user will see or need.

WriteSurf.c

If the solvent-accessible surface area is calculated, a file containing all of the results for each amino acid is printed.

Cleanup.c

This function clears the memory.

Other files

In addition to the '`.c`' source files outlined here, there are also two '`.h`' header files: `global.h` and `function.h`. `global.h` contains the definitions of all structures and global variables within the program. `function.h` contains the prototypes for all of the functions available globally.

Bibliography

[1] D. A. Dougherty. Cation-π interactions in chemistry and biology. A new view of benzene, Phe, Tyr, and Trp. *Science*, 271:163–168, 1996.

[2] J. C. Ma and D. A. Dougherty. The cation-π interaction. *Chem. Rev.*, 97:1303–1324, 1997.

[3] N. S. Scrutton and A. R. C. Raine. Cation-π bonding and amino-aromatic interactions in the biomolecular recognition of substituted ammonium ligands. *Biochem. J.*, 319:1–8, 1996.

[4] J.L. Sussman, M. Harel, F. Frolow, C. Oefner, A. Goldman, L. Toker, and I. Silman. Atomic structure of acetylcholinesterase from *Torpedo californica*: A prototypic acetylcholine-binding protein. *Science*, 253:872–879, 1991.

[5] W. Zhong, J. P. Gallivan, Y. Zhang, L. Li, H. A. Lester, and D. A. Dougherty. From *ab initio* quantum mechanics to molecular neurobiology: A cation-π binding site in the nicotinic receptor. *Proc. Natl. Acad. Sci. (USA)*, 95:12088–12093, 1998.

[6] M. F. Perutz, G. Fermi, D. J. Abraham, C. Poyart, and E. Bursaux. Hemoglobin as a receptor of drugs and peptides: X-ray studies of the stereochemistry of binding. *J. Am. Chem. Soc.*, 108:1064–1078, 1986.

[7] M. Levitt and M. F. Perutz. Aromatic rings act as hydrogen bond acceptors. *J. Mol. Biol.*, 201:751–754, 1988.

[8] M. F. Perutz. The role of aromatic rings as hydrogen-bond acceptors in molecular recognition. *Phil. Trans. R. Soc. A*, 345:105–112, 1993.

[9] S.K. Burley and G.A. Petsko. Amino-aromatic interactions in proteins. *FEBS Lett.*, 203:139–143, 1986.

[10] C.A. Deakyne and M. Meot-Ner (Mautner). Unconventional ionic hydrogen bonds. 2. NH^+ π. complexes of onium ions with olefins and benzene derivatives. *J. Am. Chem. Soc.*, 107:474–479, 1985.

[11] D.A. Rodham, S. Suzuki, R.D. Suenram, F.J. Lovas, S. Dasgupta, W.A. Goddard III, and G.A. Blake. Hydrogen bonding in the benzene-ammonia dimer. *Nature*, 362:735–737, 1993.

[12] J. Singh and J. M. Thornton. Sirius. An automated method for the analysis of the preferred packing arrangements between protein groups. *J. Mol. Biol.*, 211:595–615, 1990.

[13] J. B. O. Mitchell, C. L. Nandi, S. Ali, I. K. McDonald, J. M. Thornton, S. L. Price, and J. Singh. Amino/aromatic interactions. *Nature*, 366:413, 1993.

[14] J. B. O. Mitchell, C. L. Nandi, I. K. McDonald, J. M. Thornton, and S. L. Price. Amino/aromatic interactions in proteins: Is the evidence stacked against hydrogen bonding? *J. Mol. Biol.*, 239:315–331, 1994.

[15] C. L. Nandi, J. Singh, and J. M. Thornton. Atomic environments of arginine side-chains in proteins. *Protein Eng.*, 6:247–259, 1993.

[16] J. B. O. Mitchell, R. A. Laskowski, and J. M. Thornton. Non-randomness in side-chain packing: The distribution of interplanar angles. *Proteins*, 29:370–380, 1997.

[17] J. Singh and J. M. Thornton. *Atlas of Protein Side-Chain Interactions*, volume 1 and 2. IRL Press, Oxford, 1992.

[18] M. M. Flocco and S. L. Mowbray. Planar stacking interactions of arginine and aromatic side-chains in proteins. *J. Mol. Biol.*, 235:709–717, 1994.

[19] M. Hendlich. Database for protein-ligand complexes. *Acta Cryst.*, D54:1178–1182, 1998.

[20] J. Wouters. Cation-π (Na^+-Trp) interactions in crystal structure of tetragonal lysozyme. *Protein Sci.*, 7:2472–2475, 1998.

[21] A. Y. Ting, I. Shin, C. Lucero, and P. G. Schultz. Energetic analysis of an engineered cation-π interaction in staphylococcal nuclease. *J. Am. Chem. Soc.*, 120:7135–7136, 1998.

[22] E. E. Abola, F. C. Bernstein, S. H. Bryant, T. F. Koetzle, and J. Weng. *Protein Data Bank.* Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester, 1987.

[23] W. L. Jorgensen. The OPLS potential functions for proteins. Energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.*, 110:1657–1666, 1988.

[24] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, 118:11225–11236, 1996.

[25] M. J. Frisch, G. W. Trucks, H. B. Schlegel, P. M. W. Gill, B. G. Johnson, M. A. Robb, J. R. Cheeseman, T. A. Keith, G. A. Petersson, J. A. Montgomery, K. Raghavachari, M. A. Al-Laham, V. G. Zakrzewski, J. V. Ortiz, J. B. Foresman, C. Y. Peng, P. A. Ayala, M. W. Wong, J. L. Andres, E. S. Replogle, R. Gomperts, R. L. Martin, D. J. Fox, J. S. Binkley, D. J. Defrees, J. Baker, J. P. Stewart, M. Head-Gordon, C. Gonzalez, and J. A. Pople. *Gaussian 94 (Revision D.3).* Gaussian, Inc., 1995.

[26] K. S. Kim, J. Y. Lee, S. J. Lee, T-K. Ha, and D. H. Kim. On binding forces between aromatic rings and quaternary ammonium compounds. *J. Am. Chem. Soc.*, 116:7399–7400, 1994.

[27] A. Pullman, G. Berthier, and R. Savinelli. Interaction of the tetramethylammonium ion with the cycles of aromatic amino acids beyond the SCF *ab initio* level. *J. Am. Chem. Soc.*, 120:8553–8554, 1998.

[28] R.A. Kumpf and D.A. Dougherty. A mechanism for ion selectivity in potassium channels: Computational studies of cation-π interactions. *Science*, 261:1708–1710, 1993.

[29] J. W. Caldwell and P. A. Kollman. Cation-π interactions: Nonadditive effects are critical in their accurate representation. *J. Am. Chem. Soc.*, 117:4177–4178, 1995.

[30] O. Donini and D. F. Weaver. Development of modified force field for cation-amino acid interactions: *Ab initio*-derived empirical correction terms with comments on cation-π interactions. *J. Comput. Chem.*, 19:1515–1525, 1998.

[31] S. Mecozzi, A. P. West Jr., and D. A. Dougherty. Cation-π interactions in simple aromatics. Electrostatics provide a predictive tool. *J. Am. Chem. Soc.*, 118:2307–2308, 1996.

[32] S. Mecozzi, A. P. West Jr., and D. A. Dougherty. Cation-π interactions in aromatics of biological and medicinal interest: Electrostatic potential surfaces as a useful qualitative guide. *Proc. Natl. Acad. Sci. USA*, 93:10566–10571, 1996.

[33] U. Hobohm, M. Scharf, R. Schneider, and C. Sander. Selection of representative protein data sets. *Protein Sci.*, 1:409–417, 1992.

[34] U. Hobohm and C. Sander. Enlarged representative set of protein structures. *Protein Sci.*, 3:522–524, 1994.

[35] P. J. Kraulis. *J. Appl. Cryst.*, 24:946–950, 1991.

[36] E. M. Duffy, P. J. Kowalczyk, and W. L. Jorgensen. Do denaturants interact with aromatic hydrocarbons in water? *J. Am. Chem. Soc.*, 115:9271–9275, 1993.

Chapter 4 Computational Studies of Cation-π Interactions and Salt Bridges

4.1 Introduction

Of the myriad noncovalent interactions that contribute to protein stability, few are both specific and strong when fully exposed to an aqueous medium. Hydrogen bond formation, for example, is opposed by competing interactions with water, and the hydrophobic effect is not specific for any particular pair of residues. Although salt bridges are specific and quite strong in non-polar media, there is debate over the extent to which water-exposed salt bridges contribute to protein stability [1–6].

Many studies estimate that a water-exposed salt bridge contributes 1-2 kcal/mol to protein stability, while others suggest a salt bridge is worth very little energetically, and may even be destabilizing. There is, however, a relatively underappreciated noncovalent binding force that is potentially both specific and strong in an aqueous environment. Recent studies have shown that cation-π interactions can be quite strong in aqueous media [7]. In addition, evidence from the preceding chapters suggests that not only are cation-π interactions widely used in protein structures, but that they are frequently found on the surfaces of proteins, exposed to aqueous solvation.

This naturally leads to the question: "Why?" In this chapter, we attempt to answer this question by comparing the attributes of cation-π interactions to those of other motifs commonly found in protein structures. In particular, we present a computational study of the relative strengths of salt bridges and cation-π interactions both in aqueous media and in a broad range of organic solvents. In doing so, we hope to define a unique role for cation-π interactions within biological systems.

While there have been other computational studies of salt bridges [8–14] and of cation-π interactions [15–20] in water, a wide range of theoretical models of differing complexities and reliabilities has been employed. Previous studies of cation-π interactions in water have relied primarily upon force field-based methods and statistical perturbation theory [15–17]. Although traditional two-body force field-based methods are able to reproduce the trends in cation-π binding abilities, they are challenged to accurately reproduce the results of both experiment and high-level it ab initio calculations. Kollman and co-workers have attempted to improve two-body de-

scriptions of the cation-π interaction by adding "10-12" functions to the Amber force field [16]. Although these functions allow the force field method to reproduce the results of high-level ab initio calculations in the gas phase, they nonetheless ignore the effects of both polarization and charge-transfer, which are important in the accurate description of cation-π interactions [21, 22]. Similarly, a previously described hybrid quantum mechanical/molecular mechanical study [18] is compromised by the use of the AM1 Hamiltonian, which is similarly poor in treating cation-π interactions.[1] Since ab initio calculations more accurately reproduce experimental results with regard to cation-π interactions [7], we anticipated that a modern implementation of aqueous solvation in a quantum mechanical calculation might be useful.

Cramer and Truhlar recently introduced the SM5.42R/HF method which combines self-consistent reaction field calculations at the Hartree-Fock level with the generalized Born approximation plus surface tension terms to calculate free energies of solvation [23–25]. The SM5.42R/HF method accurately predicts the free energies of solvation of hundreds of neutral molecules and ions, including the monomers used in this study, and has recently been used to explore solvent effects on bimolecular reactions in solution [26]. In addition, the calculations exhibit rapid convergence and proper dissociation behavior in supermolecule calculations, and they are very fast. Most importantly, the method treats the intermolecular interactions at the Hartree-Fock level, thus accounting for the effects of polarization and charge transfer [21, 22]. For these reasons, we used the SM5.42R/HF method to compare the relative strengths of cation-π interactions and salt bridges at the same level of theory, so as to determine the potential significance of the less-studied cation-π interaction vs. the well-appreciated salt bridge.

[1] See Chapter 2.

4.2 Results

4.2.1 Ab Initio Calculations

As a model for the cation-π interaction, we chose methylammonium ion binding to benzene (Figure 4.1), which mimics the interaction between lysine and phenylalanine in protein structures. For the salt bridge, we chose methylammonium ion binding to acetate, which represents the interaction between lysine and glutamate or aspartate (Figure 4.1). To calculate the interaction energies, we first optimized the geometries of the individual molecules in the gas phase using HF/6-31+G* calculations. Then, using fixed monomer geometries, we optimized the complexes in the gas phase at the HF/6-31+G* level. We find that gas phase binding energies for methylammonium to benzene and acetate are -12.5 kcal/mol and -125.5 kcal/mol respectively. Not surprisingly, the ion-pair is much stronger than the cation-π interaction in the gas phase. It should also be noted that the gas-phase geometries were restricted to exist as the ion pair. Allowing the proton to transfer to form a neutral hydrogen bonded geometry would further lower the energy [27].

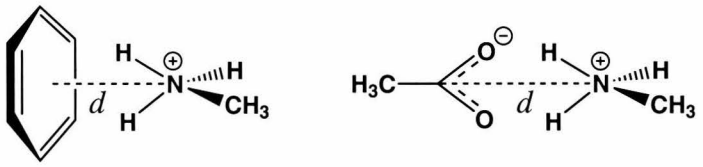


Figure 4.1: Geometries for methylammonium$\cdots$benzene (left) and methylammonium$\cdots$acetate (right). The distance,d, is reported in Figure 4.2.

To perform calculations in solution, we used the gas phase monomer geometries, altered the distances between the monomers (d in Figure 4.1), and performed SM5.42R/HF/6-31+G* supermolecule calculations to determine an "interaction energy," defined in Equation 4.1

$$\begin{aligned} Interaction\ Energy = & E_{ab(gas)} - E_{a(gas)} - E_{b(gas)} + \\ & \Delta G_{ab(sol)} - \Delta G_{a(sol)} - \Delta G_{b(sol)} \end{aligned} \tag{4.1}$$

where $E_{ab(gas)}$ is the gas phase energy of the complex, $E_{a(gas)}$ and $E_{b(gas)}$ are the gas phase energies of monomers a and b, $\Delta G_{ab(sol)}$ is the free energy of solvation of the complex, and $\Delta G_{a(sol)}$ and $\Delta G_{b(sol)}$ are the free energies of solvation of the monomers a and b. The results of these calculations with water as the solvent are shown in Figure 4.2.

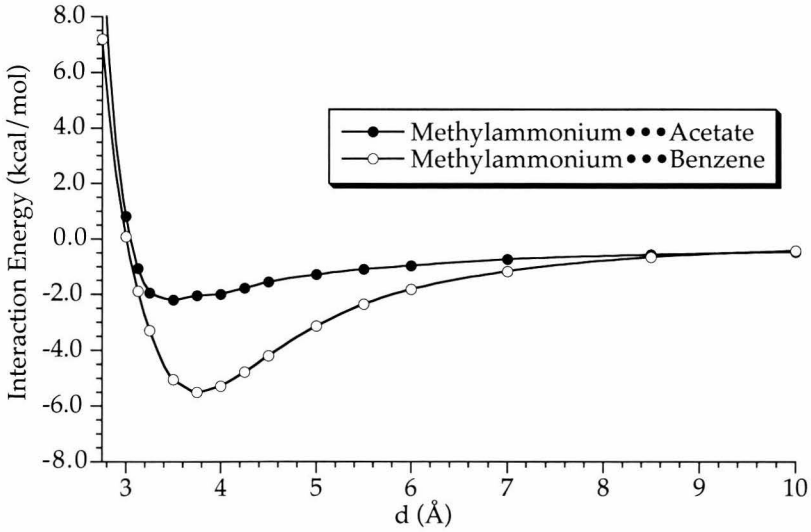


Figure 4.2: Interaction energies at the SM5.42R/HF/6-31+G* level for methylammonium$\cdots$benzene and methylammonium$\cdots$acetate in water. The data used to generate this figure can be found in Tables 4.2 and 4.3 in the **Methods** section.

Figure 4.2 clearly shows that both the salt bridge and the cation-π interaction have well-defined minima in water at this level of theory. The salt bridge experiences a broad, shallow minimum of -2.2 kcal/mol at a distance of 3.5 Å. This result is compatible with other studies, and the overall interaction surface is smooth and well-behaved, further supporting the reliability of the SM5.42R method as applied here. In contrast, the cation-π interaction experiences a substantially deeper minimum, with -5.5 kcal/mol of interaction energy at a distance of 3.75 Å. We note that Hartree-Fock theory underestimates the magnitude of the methylammonium$\cdots$benzene interaction energy in the gas phase. In particular, the level of theory used here (HF/6-31+G*)

gives a value of -12.5 kcal/mol, while the experimental ΔH is -18.8 kcal/mol [28]. As such, it may well be that the approach used here underestimates the cation-π well depth in water. While higher levels of theory perform better on the gas phase binding energy (e.g., the MP2/6-31+G* value is -17.3 kcal/mol), such methods are not presently compatible with the SM5.42R method. Since any change to a higher level of theory is expected to increase the magnitude of the cation-π interaction in water, our basic conclusion that the cation-π interaction is stronger than a typical salt bridge in water remains.

What is perhaps most striking is that in aqueous solution, the magnitude of the salt bridge interaction is reduced by over fifty-fold relative to the gas phase, whereas the cation-π interaction is reduced by less than three-fold. This suggests that cation-π interactions will remain strong across a wide range of dielectric constants. To test this notion, we performed calculations analogous to those shown in Figure 4.2 in differing solvents. The results of calculations using carbon tetrachloride, ethyl acetate, ethanol, and acetonitrile as solvents are shown in Table 4.1.

solvent	dielectric	cation-π interaction energy	salt bridge interaction energy
none	1.0	-12.5	-125.5
CCl_4	2.23	-7.8	-53.4
$CH_3CO_2CH_2CH_3$	5.99	-6.2	-19.7
CH_3CH_2OH	24.85	-5.6	-5.2
CH_3CN	37.5	-5.6	-3.8
H_2O	78.0	-5.5	-2.2

Table 4.1: Interaction energies (kcal/mol) at the SM5.42R/HF/6-31+G* level for methylammonium$\cdots$benzene and methylammonium$\cdots$acetate in a range of solvents.

As might be anticipated, both the cation-π interaction and the salt bridge show slightly deeper minima in acetonitrile than in water. In the non-polar solvents carbon tetrachloride and ethyl acetate, the results tend to approach the gas phase numbers, and the energy minimum of the salt bridge is substantially deeper than that of the cation-π interaction. Interestingly, the minima for both the cation-π interaction and the salt bridge are very similar in ethanol. A plot of the dielectric dependence of the

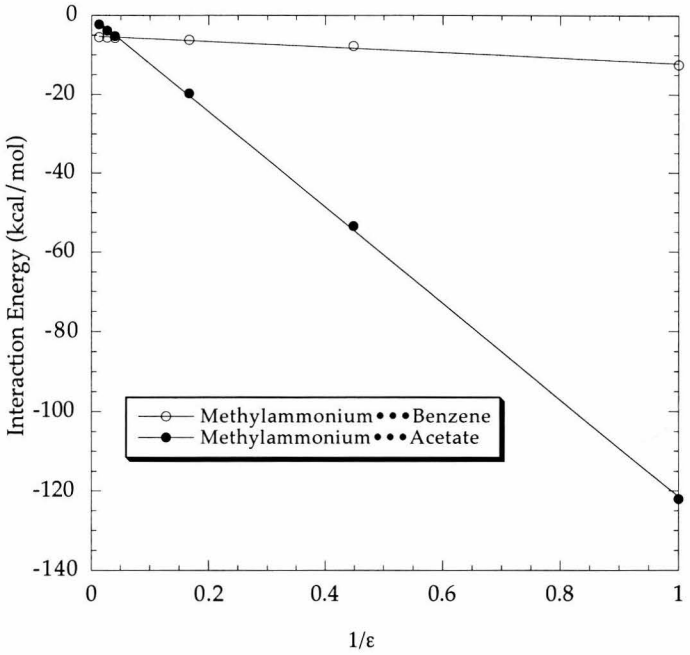


Figure 4.3: Interaction energies at the SM5.42R/HF/6-31+G* level for methylammonium$\cdots$benzene and methylammonium$\cdots$acetate in water, acetonitrile, ethanol, carbon tetrachloride, and the gas phase, plotted against the dielectric constant of the medium. $R > 0.99$ for both lines. The intersection of the lines is at $\varepsilon = 22.7$.

binding energies is shown in Figure 4.3. Clearly variations in the dielectric constant of the medium are the overwhelming contributor to variations in the strength of the interactions. This is to be is expected because of the dependence of the Born solvation model as implemented here on the dielectric constant of the medium. Nonetheless, Figure 4.3 predicts that the strengths of the cation-π interaction and the salt bridge are approximately equal in a solvent with a dielectric constant of 22.7 – slightly less polar than ethanol, and slightly more polar than acetone (ε=20.7).

Given these results, it is perhaps tempting to suggest that the minima in Figure 4.2 and Table 4.1 may represent ΔG for the association of the two partners.[2] This would suggest that these interactions have association constants in the range observable by methods such as NMR. However, to obtain a better estimate of the association constant of the binding process, one should use Equation 4.2 to integrate the potential

[2]Jorgensen has suggested that this interaction energy approximates ΔG_{bind} [29].

of mean force (PMF) curves of Figure 4.2 [16].

$$K_a = 4\pi \int_0^{r_{cut}} r^2 e^{-\frac{w(r)}{kT}} dr \tag{4.2}$$

This approach has shortcomings because the curves above were derived using only a very small number of configurations, rather than by sampling the large number of configurations typically examined in a statistical perturbation study. Because only a limited number of configurations was sampled, any estimate of the association constant will be too favorable, and will thus represent an upper limit to the magnitude of the interaction. The calculated association constant for the cation-π interaction between methylammonium and benzene in water is 401 M^{-1}, while that for the salt bridge between methylammonium and acetate is 6 M^{-1}. This suggests that the upper limit for ΔG_{bind} for the cation-π interaction is approximately -3.6 kcal/mol, while that of the salt bridge is approximately -1.0 kcal/mol. We again stress that these values must be considered with caution because of the limited number of configurations and the symmetry constraints imposed. For example, in a study of a similar system using molecular mechanics methods, Kollman found that relaxing constraints similar to those employed here diminished the binding constant by a factor of 30 [16]. Nonetheless, we do note that these numbers are consistent with earlier computational studies that employed similar constraints. In particular, the value of ΔG_{bind} for the cation-π interaction (-3.6 kcal/mol) agrees remarkably well with the value of -3.2 kcal/mol determined by Kollman and co-workers for a similarly constrained simulation [16]. In addition, the value for ΔG_{bind} for the salt bridge is in good agreement with experimental values ranging from -0.2 to -1.25 kcal/mol for surface exposed salt bridges determined by Fersht and co-workers [4]. This agreement with both experiment and theory further demonstrates the applicability of the SM5.42R/HF method for calculating weak interactions in solution.

We propose the following simple explanation for why the formation of a cation-π interaction may be favored over the formation of a salt bridge in water. To form an ion pair in water, two well-solvated ions must become desolvated to a considerable

extent. The binding energy of the complex may be very strong, but the desolvation penalty may be of comparable magnitude. In addition, although the ion pair complex is strongly polar, to some extent there is a neutralization of charge in bringing two opposite charges in close contact. This will diminish the long range Born solvation of the ion pair relative to the isolated ions. To form a cation-π interaction, a cation must pay a similar penalty to partially desolvate. However, the aromatic partner is poorly solvated by water, and so the desolvation associated with cation binding is most likely beneficial. In addition, no "charge neutralization" occurs, and so the cationic complex can still reap the substantial benefits of long-range solvation by water.

4.2.2 Protein Studies

The possibility that cation-π interactions are strong and specific in aqueous media suggests that exposed cation-π interactions may also contribute to protein structure and stability. In Chapter 3, we demonstrated that cation-π interactions are quite common within proteins. In that chapter, it was suggested, though not proven, that cation-π interactions tend to occur on the surfaces of proteins, rather than buried in the cores. To more rigorously explore this hypothesis in the context of the present chapter, we have determined the extent to which cation-π interactions are exposed to water in protein structures. To do this, we implemented within the CaPTURE (Cation-π Trends Using Realistic Electrostatics) program a variant of the GEPOL method [30] to calculate the solvent accessible surface area (SASA) of individual interactions. Traditionally, calculation of relative surface accessibility of amino acids has relied on comparing the SASA of an individual amino acid within the protein to the SASA of the same amino acid in a GXG peptide (where G is glycine and X is the amino acid). However, in a cation-π interaction, since the interacting partners are in contact with one another, their surface exposed area is by definition less than that of the amino acid in a GXG peptide, even though the interacting pair *as a unit* may be highly exposed. To overcome this difficulty, we have chosen to calculate the

accessibility of a cation-π interaction considered as a unit within the protein, relative to the accessibility of the amino acids in their identical orientation, but not in the context of the protein. To do this, the solvent accessible surface area for each partner in the cation-π interaction is calculated. The total exposed area of the two partners is then compared to the maximum exposed area of the two partners. This is determined by "excising" the interaction from the rest of the protein. The coordinates of the amino acids involved in the cation-π interaction, as well as the backbone coordinates of the $(i-1)$ and $(i+1)$ residues of both the cation and the aromatic are retained, and the rest of the atoms in the protein are deleted. The solvent accessible surface area of the cation-π interaction in this minimal model is calculated. This represents the maximum possible surface area of the cation-π interaction taken *as a unit*. The ratio of the exposed area of the cation-π interaction within the protein to the maximum area of the cation-π interaction considered by itself can be expressed as a "percent exposed". Using this method, we can also determine the "percent exposed" value for each amino acid within the dataset.

As a dataset of proteins, we chose the 593 non-homologous proteins evaluated in Chapter 3. The exposure of each amino acid, as well as 2878 energetically significant cation-π interactions, was calculated using the method outlined. The results are shown in Figure 4.4.

The results in Figure 4.4 are largely as expected for both the cationic and aromatic amino acids. Phe, Tyr, and Trp tend to be buried within the proteins with over 40% of the aromatics completely buried. In addition, the aromatic amino acids are rarely highly exposed – only 20% have more than 20% of their surfaces exposed to water. Not surprisingly, the cationic amino acids Lys and Arg tend to be exposed – 70% of all cationic amino acids expose more than 20% of their surfaces to water, and much less than 10% of these residues are completely buried. It might be anticipated that cation-π interactions would resemble the average of the cationic and aromatic amino acid curves in Figure 4.4. However, this is not the case. Instead, cation-π interactions tend to behave as a hybrid between the cationic and aromatic amino acids. On the one hand, cation-π interactions are rarely completely buried within proteins. Only 11% of

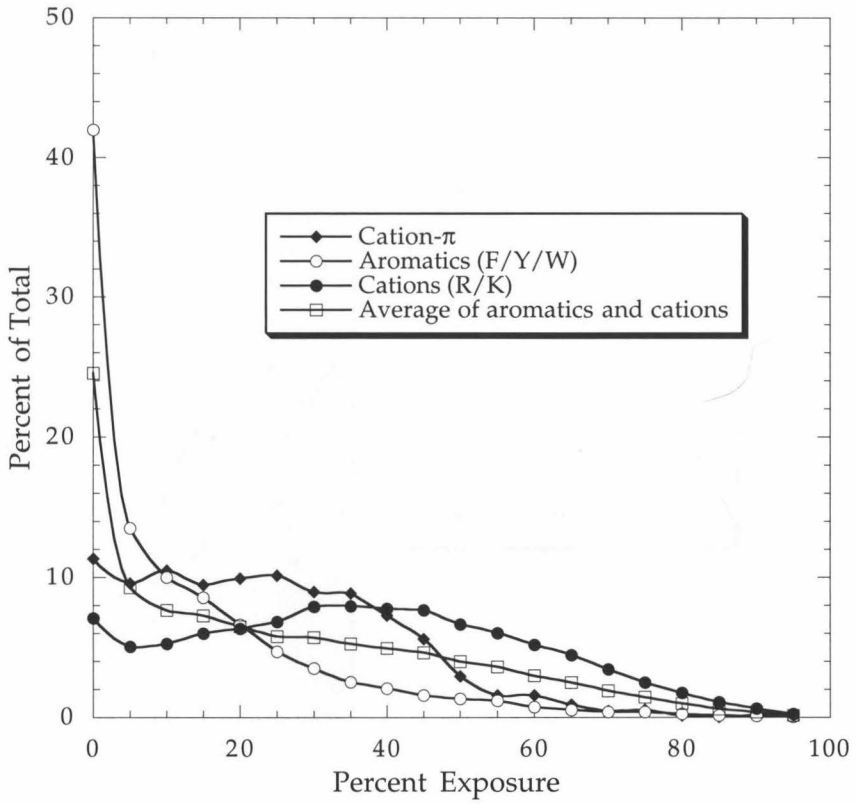


Figure 4.4: The percentage of the total number of residues (interactions) that have a given relative exposure. Abbreviations: F=phenylalanine, Y=tyrosine, W=tryptophan, R=arginine, K=lysine.

cation-π interactions are completely buried, compared to 7% for cationic amino acids and 42% for the aromatics. In this regard, they tend to behave more like cations than like aromatics. On the other hand, while cation-π interactions are more often highly exposed than are aromatics, they are certainly less often exposed than are the cationic amino acids. This is perhaps not surprising as a cation-π interaction is by definition more hydrophobic than a cationic amino acid.

Nevertheless, the question of whether cation-π interactions are "exposed" still remains. The choice of what degree of exposure constitutes "exposed" is somewhat arbitrary. Previous workers have offered varying criteria for exposure. In particular, Rost and Sander have suggested several models [31]. The first is a two-state model in which amino acids whose exposed area is greater than 16% are considered exposed,

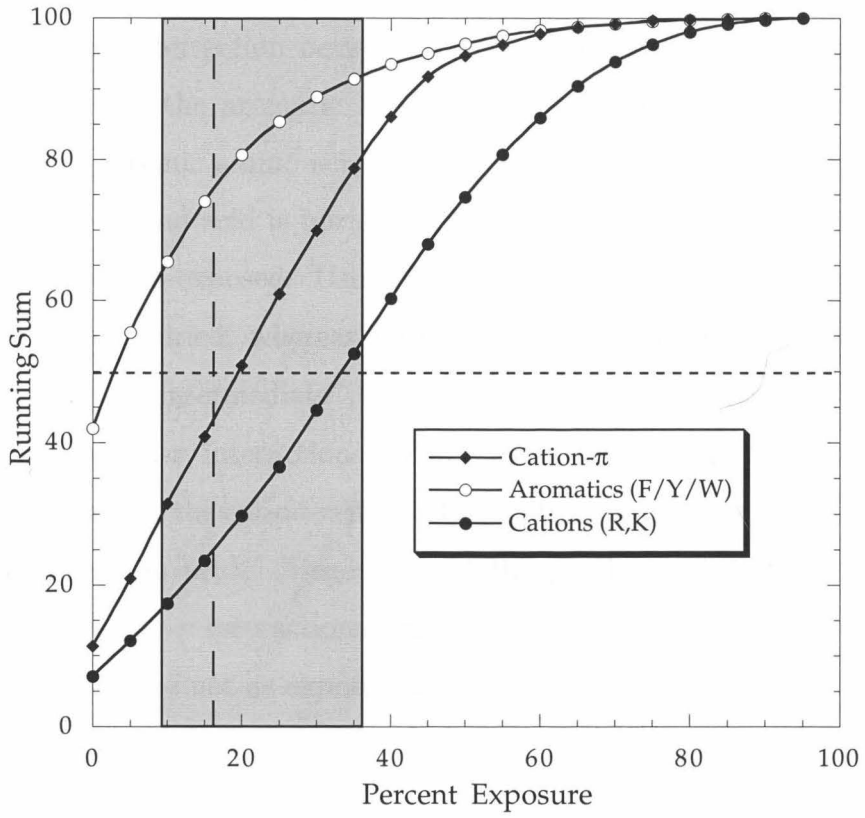


Figure 4.5: Integration of the data in Figure 4.4 for cationic amino acids, aromatic amino acids, and cation-π interactions. The vertical dotted line represents the 16% criterion in the two-state model. The shaded box represents the "intermediate" state of the three-state model (9% < exposure < 36%).

while those less than 16% are considered buried. The second is a three-state model in which amino acids whose exposed area is < 9% are considered buried, those between 9 and 36% exposure are intermediate, and those > 36% are considered exposed. Figure 4.5 integrates the data for cationic amino acids, aromatic amino acids, and cation-π interactions shown in Figure 4.4.

The vertical dotted line represents the 16% criterion in the two-state model – points to the right of this line are exposed, points to the left are buried. The shaded box represents the "intermediate" state of the three-state model (9% < exposure < 36%). Points to the left of the box are considered buried and those to the right are considered exposed in the three-state model. To determine the nature of a particular interaction, we define the PE_{50} as the percent exposure at which 50% of the amino

acids are more exposed than this point and 50% are less exposed. For a particular curve, PE_{50} is the intersection between it and the horizontal dotted line in Figure 4.5. PE_{50} is 3% for the aromatic amino acids, 20% for the cation-π interactions and 34% for the cationic amino acids. Using the two-state model suggests that the average aromatic amino acid is buried, and that the average cation-π interaction or cationic amino acid is exposed. Using the three-state model, aromatic amino acids are still considered buried, whereas both cation-π interactions and cationic amino acids are considered "intermediate", although the cationic amino acids have a higher PE_{50} than the cation-π interactions. Thus, using two different sets of criteria for determining exposure, the cation-π interaction falls into the same broad classification as the cationic amino acids. Regardless of the particular (arbitrary) cutoffs, the trend is clear – cation-π interactions tend not to be buried, reflecting their cationic character, but they are not as exposed as cationic residues by themselves, reflecting their aromatic component.

4.3 Summary

The results presented in this chapter represent the state-of-the-art in continuum calculations and compare, at a consistent level of theory, the strengths of salt bridges and cation-π interactions in water and in a range of organic solvents. The SM5.42R/HF method is well behaved for this problem, and, while future efforts may refine the present results, we believe the basic trend will survive. In particular, we suggest that in water, a cation-π interaction is more stabilizing than an analogous salt bridge. Not surprisingly, the strengths of the interactions are strongly dependent on the dielectric constant of the medium, as would be expected for interactions with large electrostatic components. Interestingly, the cation-π interaction maintains its strength across a range of solvents. Whereas the strength of the salt bridge is reduced over 50-fold on moving from the gas phase to water, the strength of a cation-π interaction is weakened by less than a factor of three. Thus, cation-π interactions can provide strong, specific interactions on the surfaces of proteins. Consistent with this prediction, we find that

cation-π interactions are rarely buried within proteins in the Protein Data Bank, and more often than not, are exposed to water. These results suggest a novel mechanism to increase the stability of proteins. Most efforts at protein design have emphasized optimization of packing in the core of the protein [32]. Engineering solvent-exposed cation-π interactions, especially at the interfaces between domains or subunits, may provide a complementary strategy.

Perhaps most importantly, these results may suggest the *raison d'être* for cation-π interactions in protein structures. In contrast to interactions such as salt bridges and hydrogen bonds, which are severely attenuated in water, or the hydrophobic effect, which lacks specificity, cation-π interactions are both strong and specific in water. As such, cation-π interactions may play a unique role by providing a means to stabilize regions of proteins that are exposed to solvent.

4.4 Appendix 1 – Methods

All calculations were performed using Gaussian 94, Rev. D3 [33], enhanced with the Minnesota Gaussian Solvation Module Version 98.2.5 (GSM) [23]. All solution phase calculations were performed with SCRF (self-consistent reaction field) model 2, using the option ISCRF=2. For the $CH_3NH_3^+ \cdots$ benzene complex, the nitrogen was constrained to the C_6 axis of benzene, the complex had C_s symmetry, and the monomer geometries were fixed at their gas phase minima. For the $CH_3NH_3^+ \cdots$ acetate complex, the distance between the N and the carboxylate carbon was varied, keeping all other parameters fixed. The solution phase minimum energy geometry was a "doubly-hydrogen bonded" structure that placed the carboxylate oxygens nearly coplanar with the ammonium hydrogens as shown in Figure 4.1. Consideration of other geometries gave comparable results. For solution phase calculations, supermolecule calculations with the SM5.42R/HF/6-31+G* method produced an "interaction energy," defined in Equation 4.1.

To validate the performance of the method, we compared the calculated $\Delta G_{(sol)}$ for the monomers methylammonium, benzene and acetate to the experimentally de-

termined values in water. In all cases, these agree well with experiment. For acetate, $\Delta G_{(sol)}$ calcd. = -77 kcal/mol, $\Delta G_{(sol)}$ expt. = -77 kcal/mol; methylammonium $\Delta G_{(sol)}$ calcd. = -76 kcal/mol, $\Delta G_{(sol)}$ expt. = -73 kcal/mol; benzene $\Delta G_{(sol)}$ calcd. = -0.6 kcal/mol, $\Delta G_{(sol)}$ expt. = -0.9 kcal/mol [23, 34]. Anilinium, a cation comparable in size to the cation-π complex of Figure 4.1, is also well treated by the method ($\Delta G_{(sol)}$ calcd. = -68 kcal/mol, $\Delta G_{(sol)}$ expt. = -68 kcal/mol). Since the method treats all the individual molecules properly, as well as a larger molecule that contains the relevant functional groups, it is reasonable to expect good performance for the supermolecules of Figure 4.1. The largest discrepancy between experiment and theory involves methylammonium ion, although it should be noted that the error in *experimental* values for ions are often several kcal/mol. In this study, methylammonium is used in calculations for both the cation-π interaction and the salt bridge. Thus, to the extent that this value may differ from experiment, it will effect both calculations in the same way.

Tables 4.2 and 4.3 contain the results used to generate Figure 4.2. Inspection of $\Delta G_{(sol)}$ for both complexes is quite revealing. For the salt bridge, $\Delta G_{(sol)}$ is continuosly reduced as the partners approach one another. For the cation-π interaction, $\Delta G_{(sol)}$ remains relatively constant – consistent with the cationic nature of the interaction.

Calculations on protein structures were performed using the CaPTURE program (Cation-π Trends Using Realistic Electrostatics)[3], modified to use the GEPOL algorithm [30] to determine the solvent accessible surface area of individual amino acids and pairs. A solvent radius of 1.4 Å was used for the probe and van der Waals radii were taken from the work of Rose *et al* [35].

[3]CaPTURE is described in Chapter 3.

distance (Å)	$E_{(gas)}$ (hartrees)	$E_{(sol)}$ (hartrees)	$\Delta G_{(sol)}$ (kcal/mol)
2.75	–323.0130812	–323.0519352	–24.381
3.0	–323.0165081	–323.0670636	–31.724
3.125	–323.0134842	–323.0700798	–35.514
3.25	–323.0089197	–323.0714869	–39.262
3.5	–322.9977698	–323.0718733	–46.501
3.75	–322.9861580	–323.0716323	–53.636
4.0	–322.9752304	–323.0715380	–60.434
4.25	–322.9653184	–323.0711991	–66.441
4.5	–322.9564439	–323.0708662	–71.801
5.0	–322.9414938	–323.0704304	–80.909
5.5	–322.9295546	–323.0701273	–88.211
6.0	–322.9198377	–323.0699184	–94.177
7.0	–322.9049581	–323.0695550	–103.286
8.5	–322.8896950	–323.0692843	–112.694
10.0	–322.8792790	–323.0691178	–119.126
$CH_3NH_3^+$	–95.57415921	–95.69577654	–76.316
$CH_3CO_2^-$	–227.2495112	–227.3725958	–77.237

Table 4.2: $E_{(gas)}$, $E_{(sol)}$, and $\Delta G_{(sol)}$, for the $CH_3NH_3^+ \cdots$acetate complex calculated at the SM5.42R/HF/6-31+G* level. $\Delta G_{(sol)}$ is the energy difference between $E_{(gas)}$ and $E_{(sol)}$.

distance (Å)	$E_{(gas)}$ (hartrees)	$E_{(sol)}$ (hartrees)	$\Delta G_{(sol)}$ (kcal/mol)
2.75	–326.2974629	–326.3963187	–62.033
3.0	–326.3040451	–326.4076982	–65.043
3.125	–326.3049767	–326.4108134	–66.414
3.25	–326.3050329	–326.4130617	–67.789
3.5	–326.3037308	–326.4158784	–70.374
3.75	–326.3016819	–326.4165876	–72.104
4.0	–326.2995494	–326.4162292	–73.218
4.25	–326.2975917	–326.4154505	–73.958
4.5	–326.2958886	–326.4145117	–74.437
5.0	–326.2932258	–326.4128132	–75.042
5.5	–326.2913382	–326.4115623	–75.442
6.0	–326.2899857	–326.4107064	–75.753
7.0	–326.2882708	–326.4096779	–76.184
8.5	–326.2869585	–326.4088708	–76.501
10.0	–326.2863071	–326.4084967	–76.675
20.0	–326.2853837	–326.4079465	–76.909
C_6H_6	–230.7110925	–230.712035	–0.591

Table 4.3: $E_{(gas)}$, $E_{(sol)}$, and $\Delta G_{(sol)}$, for the $CH_3NH_3^+ \cdots$ benzene complex calculated at the SM5.42R/HF/6-31+G* level. $\Delta G_{(sol)}$ is the energy difference between $E_{(gas)}$ and $E_{(sol)}$.

4.5 Appendix 2 – Sample Gaussian Input Files

The following are sample input files for Gaussian 94 for the $CH_3NH_3^+ \cdots$ benzene and $CH_3NH_3^+ \cdots$ acetate complexes. Calculations at different distances were performed by allowing the d variable to change.

4.5.1 $CH_3NH_3^+ \cdots$ Benzene Z-matrix

```
$RunGauss
%Mem=32MB
#p rhf/6-31+g* scf=(direct,tight,novaracc) pop=min test

CH3NH3+ benzene 6-31+g* gsm d=3.75

1,1
X
X 1 1.0
X 2 1.0 1 90.0
C 1 r1 2 90.0 3 0.0
C 1 r1 2 90.0 3 60.0
C 1 r1 2 90.0 3 120.0
C 1 r1 2 90.0 3 180.0
C 1 r1 2 90.0 3 240.0
C 1 r1 2 90.0 3 300.0
X 4 1.0 1 90.0 2 0.0
X 5 1.0 1 90.0 2 0.0
X 6 1.0 1 90.0 2 0.0
X 7 1.0 1 90.0 2 0.0
X 8 1.0 1 90.0 2 0.0
X 9 1.0 1 90.0 2 0.0
H 4 r2 10 ach 2 180.0
H 5 r2 11 ach 2 180.0
H 6 r2 12 ach 2 180.0
H 7 r2 13 ach 2 180.0
H 8 r2 14 ach 2 180.0
H 9 r2 15 ach 2 180.0
N 1 d 4 90.0 3 0.0
C 22 rnc 1 ang 4 180.0
H 23 rch1 22 cha1 1 0.0
H 22 rnh1 23 nha1 24 180.0
H 23 rch2 22 cha2 25 chd
H 23 rch2 22 cha2 25 -chd
H 22 rhn2 23 nha2 24 nhd
H 22 rhn2 23 nha2 24 -nhd
```

```
d=3.75
ang=129.95970693
rnc=1.50695973
rch1=1.07800176
cha1=108.14397235
rnh1=1.01126514
nha1=111.63851485
rch2=1.07807961
cha2=108.13096365
chd=59.99795508
rhn2=1.01126634
nha2=111.62792752
nhd=59.99832444
r1=1.38809164
r2=1.07566132
ach=90.

$CM2 ISCRF=2 ICDS=8 IAQU=1 ICMD=8 $END
```

4.5.2 $CH_3NH_3^+ \cdots$ Acetate Z-matrix

```
$RunGauss
%Mem=32MB
#p rhf/6-31+g* scf=(direct,tight) pop=min test

methyl-ammonium/acetate 6-31+g* gsm d=3.75

0 1
N
C 1 rcn
H 2 rch1 1 rha1
H 1 rnh1 2 rna1 3 180.0
H 2 rch2 1 rha2 4 nhdih
H 2 rch2 1 rha2 4 -nhdih
H 1 rnh2 2 rna2 3 chdih
H 1 rnh2 2 rna2 3 -chdih
C 1 d 4 ang1 3 dih1
O 9 rco1 1 ang2 3 dih2
C 9 rcc 10 ccoa1 3 dih3
H 11 arh1 10 arha1 9 0.0
H 11 arh2 10 arha2 9 rhd
H 11 arh2 10 arha2 9 -rhd
O 9 rco2 11 occa1 12 0.0

ang1=54.53975247
```

```
ang2=64.41786955
d=3.75
dih1=222.92249489
dih2=39.98401842
dih3=170.87183184
rcn=1.50715703
rch1=1.07809806
rch2=1.07809841
rnh1=1.01122464
rnh2=1.01122615
rha1=108.14559
rha2=108.14503852
rna1=111.6358302
rna2=111.63544777
nhdih=59.99984654
chdih=60.00004857
rco1=1.23873918
rco2=1.23715177
rcc=1.54738322
ccoa1=114.91243311
arh1=1.08530293
arha1=140.30124156
arh2=1.08795034
arha2=93.43710349
rhd=126.35484655
occa1=116.33222371

$CM2 ISCRF=2 ICDS=8 IAQU=1 ICMD=8 $END
```

Bibliography

[1] A. Matouschek, J. T. Kellis Jr., L. Serrano, M. Bycroft, and A. R. Fersht. Transient folding intermediates characterized by protein engineering. *Nature*, 346:440–445, 1990.

[2] L. R. Brown, A. De Marco, R. Richarz, G. Wagner, and K. Wüthrich. The influence of a single salt bridge on static and dynamic features of the globular solution conformation of the basic pancreatic trypsin inhibitor. *J. Biochem.*, 88:87–95, 1978.

[3] S. Dao-pin, U. Sauer, H. Nicholson, and B. W. Matthews. Contributions of engineered surface salt bridges to the stability of T4 lysozyme determined by directed mutagenesis. *Biochemistry*, 30:7142–7153, 1991.

[4] A. Horovitz, L. Serrano, B. Avron, M. Bycroft, and A. R. Fersht. Strength and co-operativity of contributions of surface salt bridges to protein stability. *J. Mol. Biol.*, 216:1031–1044, 1990.

[5] H.-J. Schneider, T. Schiestel, and P. Zimmermann. The incremental approach to noncovalent interactions: Coulomb and van der Waals effects in organic ion pairs. *J. Am. Chem. Soc.*, 114:7698–7703, 1992.

[6] J. P. Schneider, J. D. Lear, and W. F. DeGrado. A designed buried salt bridge in a heterodimeric coiled coil. *J. Am. Chem. Soc.*, 119:5742–5743, 1997.

[7] J. C. Ma and D. A. Dougherty. The cation-π interaction. *Chem. Rev.*, 97:1303–1324, 1997.

[8] K. T. No, K-Y. Nam, and H. A. Scheraga. Stability of like and oppositely charged organic ion pairs in aqueous solution. *J. Am. Chem. Soc.*, 119:12917–12922, 1997.

[9] S. Saigal and J. Pranata. Monte Carlo simulations of guanidinium acetate and methylammonium acetate ion pairs in water. *Bioorg. Chem.*, 25:11–21, 1997.

[10] X. Barril, C. Alemán, M. Orozco, and F. J. Luque. Salt bridge interactions: Stability of the ionic and neutral complexes in the gas phase, in solution, and in proteins. *Proteins*, 32:67–79, 1998.

[11] Z. S. Hendsch and B. Tidor. Do salt bridges stabilize proteins? A continuum electrostatic analysis. *Protein Sci.*, 3:211–226, 1994.

[12] B. Honig and A. Nicholls. Classical electrostatics in biology and chemistry. *Science*, 268:1144–1149, 1995.

[13] A. H. Elcock. The stability of salt bridges at high temperatures: Implications for hyperthermophilic proteins. *J. Mol. Biol.*, 284:489–502, 1998.

[14] R. Luo, L. David, H. Hung, J. Devaney, and M. K. Gilson. Strength of solvent-exposed salt-bridges. *J. Phys. Chem. B*, 103:727–736, 1999.

[15] E. M. Duffy, P. J. Kowalczyk, and W. L. Jorgensen. Do denaturants interact with aromatic hydrocarbons in water? *J. Am. Chem. Soc.*, 115:9271–9275, 1993.

[16] C. Chipot, B. Maigret, D. A. Pearlman, and P. A. Kollman. Molecular dynamics potential of mean force calculations: A study of the toluene-ammonium π-cation interactions. *J. Am. Chem. Soc.*, 118:2998–3005, 1996.

[17] R.A. Kumpf and D.A. Dougherty. A mechanism for ion selectivity in potassium channels: Computational studies of cation-π interactions. *Science*, 261:1708–1710, 1993.

[18] J. Gao, L. W. Chou, and A. Auerbach. The nature of cation-π binding: Interactions between tetramethylammonium ion and benzene in aqueous solution. *Biophys. J.*, 65:43–47, 1993.

[19] M. Gaberscek and J. Mavri. Phenol forms complexes with tetramethylammonium ions in aqueous solution? *Chem. Phys. Lett.*, 308(5-6):421–427, 1999.

[20] M. A. L. Eriksson, P.-Y. Morgantini, and P. A. Kollman. Binding of organic cations to a cyclophane host as studied with molecular dynamics simulations and free energy calculations. *J. Phys. Chem. B*, 103:4474–4480, 1999.

[21] J. W. Caldwell and P. A. Kollman. Cation-π interactions: Nonadditive effects are critical in their accurate representation. *J. Am. Chem. Soc.*, 117:4177–4178, 1995.

[22] E. Cubero, F. J. Luque, and M. Orozco. Is polarization important in cation-π interactions? *Proc. Natl. Acad. Sci. USA*, 95:5976–5980, 1998.

[23] J. Li, G. D. Hawkins, C. J. Cramer, and D. G. Truhlar. Universal reaction field model based on *ab initio* Hartree-Fock theory. *Chem. Phys. Lett.*, 288:293–298, 1998.

[24] C. J. Cramer and D. G. Truhlar. Implicit solvation models: Equilibria, structure, spectra, and dynamics. *Chem. Rev.*, 99:2161–2200, 1999.

[25] J. Li, T. Zhu, C. J. Cramer, and D. G. Truhlar. New class IV charge model for extracting accurate partial charges from wave functions. *J. Phys. Chem. A*, 102:1820–1831, 1998.

[26] Y.-Y. Chuang, M. L. Radhakrishnan, P. L. Fast, J. Cramer, and D. G. Truhlar. Direct dynamics for free radical kinetics in solution: Solvent effect on the rate constant for the reaction of methanol with atomic hydrogen. *J. Phys. Chem. A*, 103:4893–4909, 1999.

[27] Y-J. Zheng and R. L. Ornstein. What happens to salt-bridges in nonaqueous environments: Insights from quantum mechanics calculations. *J. Am. Chem. Soc.*, 118:11237–11243, 1996.

[28] C.A. Deakyne and M. Meot-Ner (Mautner). Unconventional ionic hydrogen bonds. 2. NH^+ π. complexes of onium ions with olefins and benzene derivatives. *J. Am. Chem. Soc.*, 107:474–479, 1985.

[29] M. L. Lamb and W. L. Jorgensen. Computational approaches to molecular recognition. *Curr. Opin. Chem. Biol.*, 1:449–457, 1997.

[30] E. Silla, I. Tuñón, and J. L. Pascual-Ahuir. GEPOL: An improved description of molecular surfaces II. Computing the molecular area and volume. *J. Comput. Chem.*, 12:1077–1088, 1991.

[31] B. Rost and C. Sander. Conservation and prediction of solvent accessibility in protein families. *Proteins*, 20:216–226, 1994.

[32] B.I. Dahiyat and S.L. Mayo. De novo protein design: Fully automated sequence selection. *Science*, 278(5335):82–87, 1997.

[33] M. J. Frisch, G. W. Trucks, H. B. Schlegel, P. M. W. Gill, B. G. Johnson, M. A. Robb, J. R. Cheeseman, T. A. Keith, G. A. Petersson, J. A. Montgomery, K. Raghavachari, M. A. Al-Laham, V. G. Zakrzewski, J. V. Ortiz, J. B. Foresman, C. Y. Peng, P. A. Ayala, M. W. Wong, J. L. Andres, E. S. Replogle, R. Gomperts, R. L. Martin, D. J. Fox, J. S. Binkley, D. J. Defrees, J. Baker, J. P. Stewart, M. Head-Gordon, C. Gonzalez, and J. A. Pople. *Gaussian 94 (Revision D.3)*. Gaussian, Inc., 1995.

[34] W.C. Still, A. Tempczyk, R.C. Hawley, and T. Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.*, 112:6127–6129, 1990.

[35] G. D. Rose, A. R. Geselowitz, G. J. Lesser, R. H. Lee, and M. H. Zehfus. Hydrophobicity of amino acid residues in globular proteins. *Science*, 229:834–838, 1985.

Chapter 5 Weak Electrostatic Interactions Involving Perfluorinated Aromatic Rings

5.1 Introduction

In the preceding chapters, we have emphasized the ability of aromatic rings to bind to cations via cation-π interactions. The implication has been that benzene and other aromatics act as "pseudo-anions," and thus bind cations through electrostatic interactions. If this idea is correct, one might predict that "pseudo-anions," such as benzene, might also bind "pseudo-cations," or molecules that display regions of positive electrostatic potential. Indeed this is the case. A prototype system is water$\cdots$benzene. Water binds to benzene with a calculated -1.8 kcal/mol of binding energy in the geometry shown in Figure 5.1 [1].

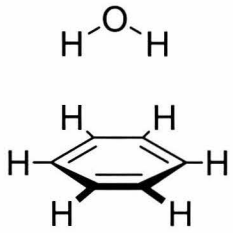


Figure 5.1: Experimentally determined geometry of water$\cdots$benzene [1].

Using standard arguments, this interaction is somewhat surprising. Water is a polar molecule, whereas benzene is generally considered to be non-polar. Furthermore, in solution, water and benzene do not mix to an appreciable extent. So why, then, are these molecules attracted to one another in the gas phase? The answer lies in the electrostatic properties of water and benzene. Although benzene does not have a dipole moment, benzene has a large, permanent quadrupole moment,[1] such that there is substantial negative electrostatic potential above and below the plane of the ring, and a belt of positive potential around the edge [2, 3]. This can be easily seen from the electrostatic potential surface of benzene, shown in Figure 5.2.

Ionic and polar species interact with this moment as anticipated from purely electrostatic arguments. Thus, inspection of the electrostatic potential surface of water, also shown in Figure 5.2, suggests that the hydrogens of water, with their region

[1] A quadrupole moment arises from an arrangement of dipoles such that there is no net dipole moment.

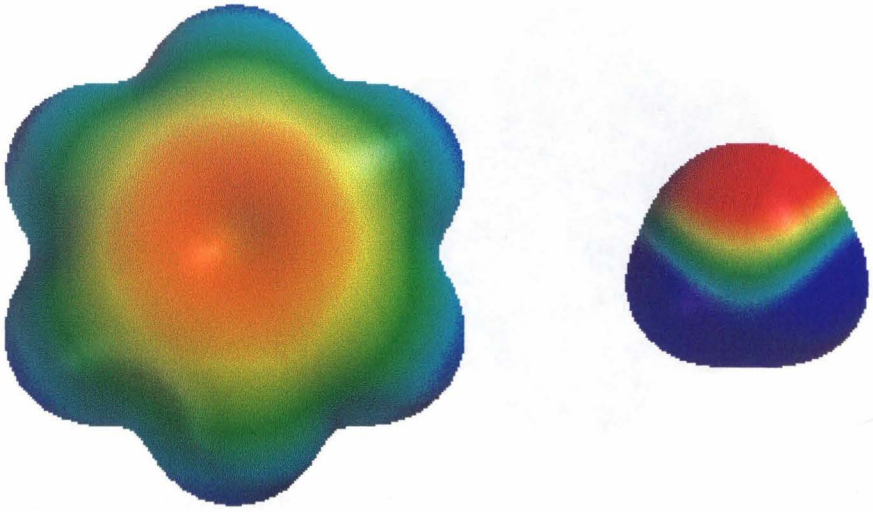


Figure 5.2: Calculated HF/6-31G** electrostatic potential surfaces of benzene (left) and water (right). *Scale: -25 – +25 kcal/mol.*

of positive electrostatic potential (shown in blue), should bind to benzene's region of negative electrostatic potential (shown in red). Indeed this is what is observed both theoretically and experimentally [1]. Thus, simple inspection of the electrostatic potential surfaces allows one to predict the likely binding geometry of the two molecules.[2]

It has long been recognized that hexafluorobenzene has a quadrupole moment that is comparable in magnitude but opposite in sign to that of benzene.[3] This effect manifests itself in the electrostatic potential surface for hexafluorobenzene shown in Figure 5.3. Relative to the electrostatic potential surface of benzene, positive regions become negative and negative positive. Inspection of the electrostatic potential surfaces of benzene (Figure 5.2) and hexafluorobenzene (Figure 5.3) suggests that benzene and hexafluorobenzene should interact favorably by binding through a face to face interaction. This stacking is observed both computationally [4] and experimentally [3], and aryl-perfluoroaryl interactions are now being employed as "supramolecular syn-

[2]This is not to imply that non-electrostatic effects are not important contributors to the binding – they certainly are. However, the geometries in these simple systems can often be predicted by consideration of electrostatics alone.

[3]The quadrupole moments of benzene and hexafluorobenzene are -29.0×10^{-40} C m^2 and $+31.7\times10^{-40}$ C m^2, respectively [3].

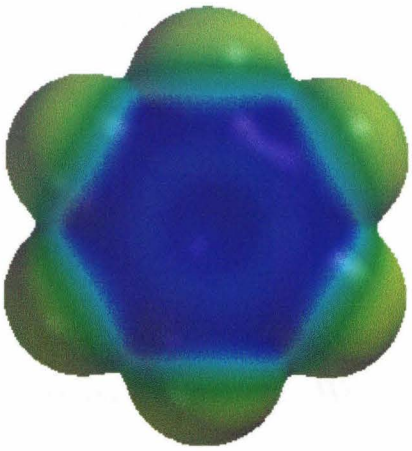


Figure 5.3: Calculated HF/6-31G** electrostatic potential surface of hexafluorobenzene. *Scale: -25 – +25 kcal/mol.*

thons" [5–7].[4]

Until this point, all interactions discussed have found experimental support. Although the water···benzene interaction may be somewhat surprising at first, consideration of the electrostatic potential surfaces of the partners explains the behavior of the system. Taking electrostatic reasoning further, one might predict that water should also bind to hexafluorobenzene, but in the geometry shown in Figure 5.4, with the water dipole moment reversed relative to the water···benzene interaction.

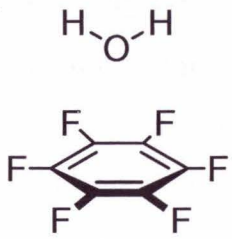


Figure 5.4: Predicted binding geometry of water···hexafluorobenzene.

While quite reasonable from an electrostatic viewpoint, from another perspective this interaction is somewhat counterintuitive. It suggests that a π-electron system should be attracted to a set of lone pair electrons. One might expect significant electron-electron repulsions that are not present for the water···benzene interaction. As such, the water···hexafluorobenzene interaction pushes the limits of electrostatic

[4]Calculations describing two of these systems are presented in an appendix to this chapter.

arguments in rationalizing noncovalent binding interactions to aromatics.

5.2 Results

5.2.1 On-Axis Geometries

To investigate whether water can bind to hexafluorobenzene, we performed ab initio molecular orbital calculations on this system. We first considered the "on-axis" complex shown in Figure 5.5.

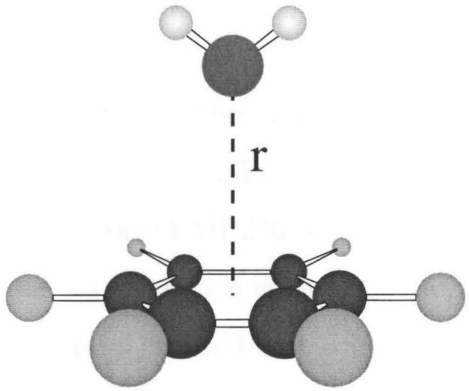


Figure 5.5: On-axis geometry of the water$\cdots$hexafluorobenzene complex.

Using Hartree-Fock theory and the 6-31G** basis set, we find that water binds to hexafluorobenzene with -2.49 kcal/mol of binding energy at an oxygen–centroid distance of 3.22 Å. Although these calculations indicate an interaction between the partners, there are two potential problems. The first is that the binding energies of complexes determined from supermolecule calculations tend to be overestimated due to basis set superposition error (BSSE). The second is that Hartree-Fock theory tends to underestimate the binding energies of neutral complexes because the effects of electron correlation are ignored. The effects of BSSE can be minimized by performing counterpoise (CP) corrections according to the method of Bernardi and Boys [8]. Furthermore, the effects of electron correlation can be accounted for in a post-Hartree-Fock method, such as Møller-Plesset (MP) perturbation theory. The results of these calculations are shown in Table 5.1.

theory	r (Å)	–binding energy (kcal/mol)
HF/6-31G**	3.22	2.49
CP-HF/6-31G**	3.50	1.67
MP2(full)/6-31G**	2.94	3.90
CP-MP2(full)/6-31G**	3.20	2.11
B3LYP/6-31G**	3.08	2.43

Table 5.1: Results of ab initio calculations for the water$\cdots$hexafluorobenzene complex shown in Figure 5.5. See Table 5.5 in Section 5.6 for the total energies.

Calculations on noncovalent systems can be challenging. Our goal here is not to present the highest level of theory imaginable for this system. Rather, we employ a competent level of theory that has been shown to handle the water$\cdots$benzene system well [1]. While more advanced levels of theory may change some quantitative aspects of our results, we are confident that no qualitative conclusions will change. The highest level of theory used here involved MP2/6-31G** calculations with counterpoise (CP) corrections for basis set superposition error (BSSE). Similar to a previous study of benzene$\cdots$hexafluorobenzene [4], we find that BSSE strongly affects the interaction distance between water and the aromatic (but no other geometrical parameters), and so this distance has been optimized with the CP correction. This level of theory is equivalent to that applied previously to water$\cdots$benzene [1], making direct comparisons meaningful. The global minimum for water$\cdots$hexafluorobenzene (Figure 5.5) is as anticipated from electrostatic arguments, and the binding energy (-2.1 kcal/mol) is quite comparable to that for water$\cdots$benzene (-1.8 kcal/mol). There appear to be no adverse electron-electron repulsions, consistent with an earlier study of polar molecules binding to hexafluorobenzene [9]. Not surprisingly, the potential energy surface is quite flat in this region, and the binding energy in the eclipsed geometry shown differs by less than 1 cal/mol from the 15° and 30° staggered geometries at the CP-MP2(full)/6-31G** level.

The magnitude of the binding energy in this system is substantial – 2.1 kcal/mol. It is perhaps most surprising given that one might expect substantial electron-electron repulsions, even though electrostatics predicts that water and hexafluoroben-

zene should bind. To estimate the fraction of the binding energy associated with electrostatics, we make the assumption that the total binding energy of the complex (E_{tot}) can be represented as the sum of the electrostatic contribution (E_{es}) and the non-electrostatic contributions (E_{non-es}). Thus, the total energy for the water$\cdots$hexafluorobenzene system can be represented as in Equation 5.1. Unfortunately, one cannot directly calculate E_{es} and E_{non-es} using ab initio calculations. One can only calculate E_{tot}. However, recall that the quadrupole moment of benzene is nearly equal in magnitude to that of hexafluorobenzene, but opposite in sign. Thus, the electrostatic contribution of water binding to benzene in the geometry of Figure 5.5 should be equal and opposite to that of water binding to hexafluorobenzene in the same geometry.

$$E_{tot\,(C_6F_6\cdots H_2O)} = E_{non-es\,(C_6F_6\cdots H_2O)} + E_{es\,(C_6F_6\cdots H_2O)} \tag{5.1}$$

$$E_{tot\,(C_6H_6\cdots H_2O)} = E_{non-es\,(C_6H_6\cdots H_2O)} + E_{es\,(C_6H_6\cdots H_2O)} \tag{5.2}$$

By making the assumption that $E_{es\,(C_6F_6\cdots H_2O)} = -E_{es\,(C_6H_6\cdots H_2O)}$ and combining Equations 5.1 and 5.2, we find that

$$E_{tot\,(C_6F_6\cdots H_2O)} + E_{tot\,(C_6H_6\cdots H_2O)} = E_{non-es\,(C_6F_6\cdots H_2O)} + E_{non-es\,(C_6H_6\cdots H_2O)} \tag{5.3}$$

Furthermore, it is reasonable to make the assumption that to first order

$$E_{non-es\,(C_6F_6\cdots H_2O)} = E_{non-es\,(C_6H_6\cdots H_2O)} \tag{5.4}$$

because the closest contacts are between the oxygen and the carbon atoms in both systems. Thus, substituting Equation 5.4 into Equation 5.3 and rearranging, we find that

$$E_{non-es\,(C_6F_6\cdots H_2O)} = E_{non-es\,(C_6H_6\cdots H_2O)} = \frac{E_{tot\,(C_6F_6\cdots H_2O)} + E_{tot\,(C_6H_6\cdots H_2O)}}{2} \tag{5.5}$$

Thus, by determining $E_{tot\,(C_6H_6\cdots H_2O)}$ using an ab initio calculation, we can estimate

$E_{non-es\,(C_6F_6\cdots H_2O)}$ from Equation 5.5 and $E_{es\,(C_6F_6\cdots H_2O)}$ using Equation 5.1.

We thus calculated $E_{tot\,(C_6H_6\cdots H_2O)}$ at the CP-MP2/6-31G** level. Surprisingly, $E_{tot\,(C_6H_6\cdots H_2O)} = +2.1$ kcal/mol — equal and opposite to $E_{tot\,(C_6F_6\cdots H_2O)}$. This implies that E_{non-es} for both systems is zero, and that the water$\cdots$hexafluorobenzene interaction is completely dominated by electrostatics. If this were true, one would predict that the energy of water binding to 1,3,5-trifluorobenzene (a molecule with a near-zero quadrupole moment, and thus limited potential for electrostatic interaction) in the geometry of Figure 5.5 would be 0. Indeed this is the case – the CP-MP2/6-31G** binding energy is +0.009 kcal/mol.

Although these results are consistent, they are at first glance surprising. Chemical intuition might suggest that E_{non-es}, which consists largely of van der Waals forces, would be favorable in the water$\cdots$hexafluorobenzene system. However, further consideration of the interaction suggests why this is not the case. If we consider two point charges of opposite sign in the gas phase, Coulomb's law predicts that they will experience the maximum attraction at the minimum distance. Using similar reasoning, water and hexafluorobenzene are attracted to one another, and become more so at smaller distances. Of course at small distances, the molecules will repel, because other non-electrostatic forces, *including van der Waals forces*, will become repulsive. This is illustrated in Figure 5.6 which shows three potential energy profiles. The red curve is the potential energy profile for an electrostatic interaction between oppositely charged partners ($-\frac{1}{r}$ dependence). The potential energy approaches 0 as r approaches ∞, and it approaches $-\infty$ as r approaches 0. In this system, the electrostatic interaction is always favorable. The blue curve represents the potential energy profile for the Lennard-Jones "12–6" potential which is commonly used to represent van der Waals forces ($\frac{1}{r^{12}} - \frac{1}{r^6}$ dependence). At very short distances, this force is quite repulsive. At a somewhat longer distance, there is a minimum, and as r approaches ∞, the red curve approaches 0. The purple curve is the sum of the red and blue curves. It, too, has a minimum — one that is shifted to shorter distances than that of the minimum of the blue curve. Clearly, at short distances, the repulsion from the "12–6" curve dominates the potential energy curve. However, because the electrostatics of the system

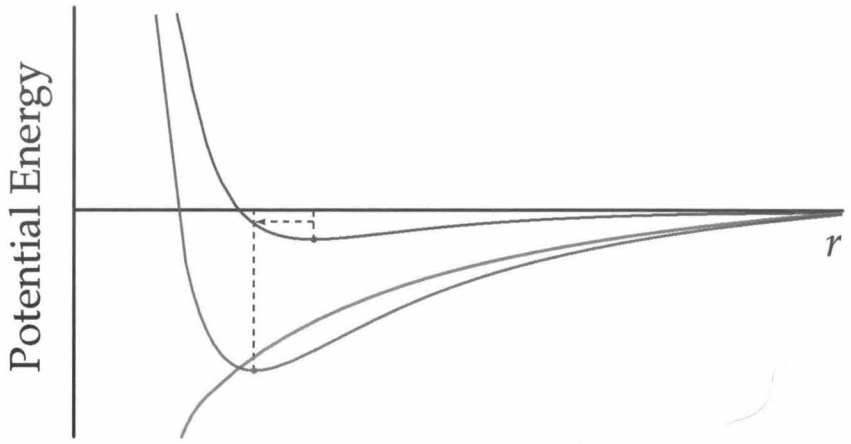


Figure 5.6: Schematic potential energy curves. Red: electrostatic interaction between oppositely charged partners ($-\frac{1}{r}$ dependence); blue: Lennard–Jones potential ($\frac{1}{r^{12}} - \frac{1}{r^6}$ dependence); purple: sum of the red and blue curves. The vertical dotted lines correspond to the minima of the two curves. The horizontal dotted line shows that at the true minimum, the Lennard–Jones potential (blue curve) is not at a minimum.

are always favorable, the minimum of the purple curve is found at a shorter distance than the minimum of the "12–6" curve. In fact, at the distance of the true minimum (represented by a purple dot), the blue curve is nearly in positive territory. This is similar to the situation observed in our calculations. Depending on the relative magnitudes of the two curves, one can envision situations where the true minimum of the purple curve corresponds to a location of the blue curve where the potential energy is ≥ 0. Similarly, in cases where the red curve is flat (such as the water$\cdots$1,3,5-trifluorobenzene interaction), the two minima will correspond. Indeed, water binds to 1,3,5-trifluorobenzene with -0.3 kcal/mol of binding energy at a distance of 3.7 Å — greater than the 3.2 Å seen in the water$\cdots$hexafluorobenzene interaction. If we consider the 0.3 kcal/mol of binding energy to be equal to the non-electrostatic component of the water$\cdots$hexafluorobenzene interaction, it suggests that electrostatics contribute 86% of the binding energy in the water$\cdots$hexafluorobenzene system. This should serve as a cautionary tale to those who attempt to decompose binding energies into their components — for interactions dominated by favorable electrostatics, van der Waals interactions serve to keep the molecules *apart*, and thus may not always be favorable.

5.2.2 Off-Axis Geometries

In addition to the on-axis geometries discussed in the previous section, we also considered other geometries for the water$\cdots$hexafluorobenzene complex. Optimization starting from an electrostatically unfavorable geometry with an O–H bond located on the C_6 axis and the hydrogen pointed toward the face of the ring led to a second minimum shown in Figure 5.7, in which the water has moved well off the C_6 axis. This structure is calculated to be only slightly less stable than the on-axis geometry

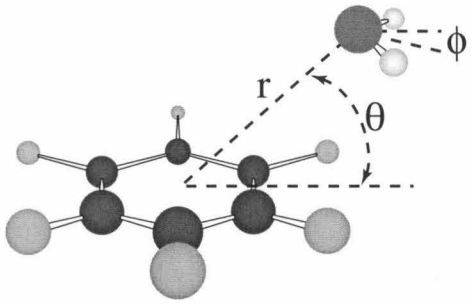


Figure 5.7: Off-axis geometry of the water$\cdots$hexafluorobenzene complex.

shown in Figure 5.5 (-1.7 vs. -2.1 kcal/mol at the CP-MP2/6-31G** level). Results at several levels of theory are presented in Table 5.2. Rather than a symmetrical

theory	r (Å)	θ	ϕ	H$\cdots$F (Å)	–binding energy (kcal/mol)
HF/6-31G**	3.96	41.8	36.9	2.61	2.95
MP2(full)/6-31G**	3.77	42.5	48.5	2.33	5.32
CP-MP2(full)/6-31G**	4.05	45.0	14.9	2.93	1.69
B3LYP/6-31G**	3.95	39.6	60.7	2.30	4.32

Table 5.2: Results of ab initio calculations for the complex shown in Figure 5.7. See Figure 5.7 for definitions of r, θ and ϕ. See Table 5.6 in Section 5.6 for the total energies.

alignment of molecular moments (dipole/quadrupole), this geometry represents a favorable alignment of local O–H and C–F bond dipoles, and so presumably still has a substantial electrostatic component. To determine whether an analogous minimum exists for water$\cdots$benzene, calculations were performed at the HF/6-31G** level. These efforts were not successful — all starting geometries led to structures similar to that previously reported for water$\cdots$benzene.

Given the increasing interest in DFT methods, we also evaluated both the on-axis and off-axis structures using the Becke3LYP density functional and the 6-31G** basis set. Since no counterpoise corrections were applied, the most appropriate comparison is to the MP2 results. In an earlier study of the benzene$\cdots$hexafluorobenzene interaction [4], DFT performed poorly, being comparable to simple Hartree-Fock calculations. However, for water$\cdots$hexafluorobenzene DFT produces good results for the global minimum, producing values that are in acceptable agreement with the MP2 results and are actually quite close to the full CP-MP2 results (Table 5.1). While CP-MP2 predicts a significant preference for the on-axis geometry, DFT finds the off-axis structure to be more stable. Both the DFT and the MP2 structures have a much larger value of ϕ, the tilt angle of the water molecule, than the CP-MP2 structure. This tilting brings the water hydrogens very close to the fluorines, well inside van der Waals contact, indicative of a hydrogen bond. Dunitz has concluded that hydrogen bonds to fluorine are not energetically significant [10], but it appears that both DFT and MP2 consider such an interaction to be important. It is clear that BSSE will be most important when noncovalent contacts to fluorine are involved, because of the large demands fluorine puts on its basis set. Thus, it is perhaps not surprising that the CP corrections are most significant for the off-axis structure.[5] For these reasons we consider the CP-MP2 level to be the best for comparisons of the two structures.

The results for the off-axis geometry suggest that a fully in-plane structure such as that shown in Figure 5.8 might be favorable. However, at the HF/6-31G** level such a structure is in fact a transition state separating two equivalent forms of the structure shown in Figure 5.7.

5.3 Summary

These results further illustrate the power of electrostatic reasoning in studies of noncovalent interactions involving aromatics. In the absence of such effects, one would never predict a stabilizing interaction for a geometry that points oxygen lone pairs

[5]There are no close contacts to fluorine in the on-axis structure.

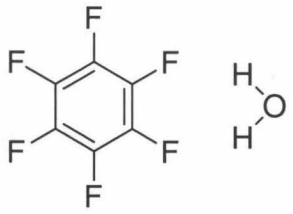


Figure 5.8: In-plane geometry of water$\cdots$hexafluorobenzene. HF/6-31G** calculations indicate that this is a transition state structure, separating two equivalent forms of the structure shown in Figure 5.7

directly into the face of a π system as observed here. As always, other effects related to polarizability and dispersion interactions certainly contribute to the binding energy, but all the structures discussed here can be accurately anticipated on the basis of electrostatics alone. These results further support the idea first put forth by Reisse [2] that aromatics such as benzene and hexafluorobenzene are correctly thought of as polar molecules, and clearly underscore the importance of electrostatic reasoning in evaluating novel intermolecular interactions.[6]

5.4 Computational Methods

All calculations were performed using Gaussian 94 Rev. D3 [12]. Geometry optimizations were first performed at the HF level using the 6-31G** basis set. Frequency calculations performed on fully optimized structures at the HF level produced no negative frequencies (except for the in-plane structure), indicating that the geometries were true minima. Using the Hartree-Fock frequency calculations, we estimate the zero-point energy (ZPE) correction to be +0.58 kcal/mol for the structure in Figure 5.5 and +0.89 kcal/mol for the structure in Figure 5.7 using the frequency scaling factor of 0.8992 suggested by Scott and Radom [13]. These results are not included in Tables 5.1 and 5.2. Counterpoise corrected optimizations (designated CP) were performed using the method of Bernardi and Boys [8]. First, the component molecules, water and hexafluorobenzene, were individually optimized at either the HF/6-31G** level or the MP2(full)/6-31G** level (the full indicates the "frozen-core" method was

[6]Recently, Danten *et al.* reported a study containing similar findings to those described here [11].

not used in the MP2 calculations – all electrons were considered). The internal coordinates of the optimized monomers were kept fixed and the complexes were optimized by scanning the potential energy surface at several points. CP corrections were performed at each point, and this corrected binding energy was minimized. The results of these calculations are designated CP-HF/6-31G** or CP-MP2(full)/6-31G**. The on-axis complex had C_{2v} symmetry, whereas the off-axis complex had C_s symmetry. Tables detailing the geometries and total energies of several of the complexes can be found in Section 5.6.

5.5 Appendix 1 – Studies of the Supramolecular Synthon Triphenylene$\cdots$ Perfluorotriphenylene

In recent years, the field of crystal engineering has seen a flourish of interest. The ability to rationally engineer solid-state structures poses great challenges. In particular, a thorough understanding of noncovalent interactions is necessary to enable one to design and manipulate solid-state architectures. Desiraju has advocated the "supramolecular synthon" approach, in which the goal of crystal engineering is to *"recognize and design synthons that are robust enough to be exchanged from one network structure to another, which ensures generality and predictability"* [14]. One such supramolecular synthon is the phenyl-perfluorophenyl system, in which the electrostatically favorable face-to-face stacking of fluorinated and non-fluorinated aromatic rings is employed to direct crystal packing. This system has been studied extensively, both theoretically [4] and experimentally [3, 5, 6, 15]. Recently, Grubbs and coworkers have extended the phenyl-perfluorophenyl supramolecular synthon to include fluorinated and non-fluorinated polycyclic aromatics [7].

To further explore such interactions, we performed ab initio calculations to investigate the energetics of the interaction between triphenylene and perfluorotriphenylene, and to compare them to the well-studied benzene$\cdots$hexafluorobenzene interaction. We first calculated the electrostatic potential surfaces of both triphenylene and perflu-

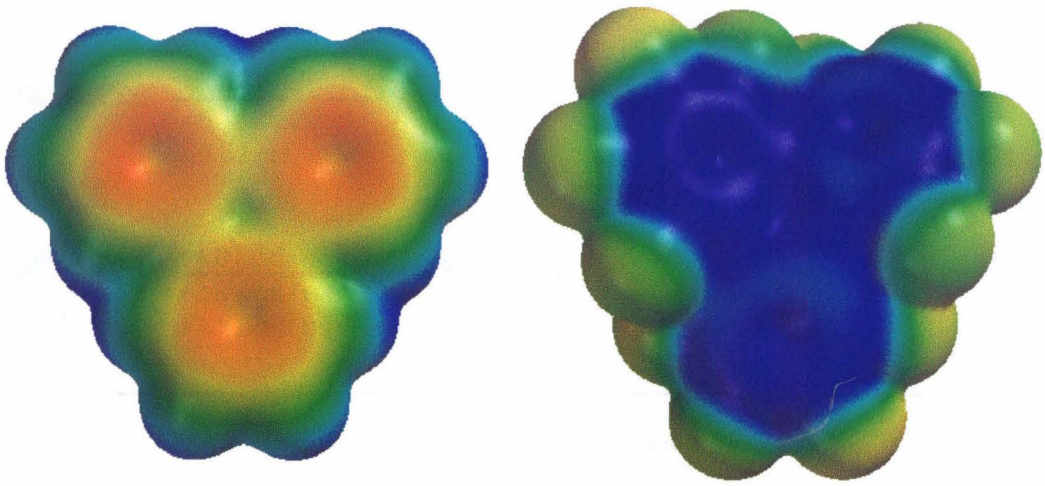


Figure 5.9: Calculated HF/6-31G** electrostatic potential surfaces of triphenylene (left) and perfluorotriphenylene (right). *Scale: -25 – +25 kcal/mol.*

orotriphenylene at the HF/6-31G** level (Figure 5.9). For triphenylene, we calculated the geometry by optimizing the structure at the HF/6-31G** level in D_{3h} symmetry. For perfluorotriphenylene, we used the coordinates from the crystal structure determined by Alex Dunn [16]. Several features are evident in Figure 5.9. The first is that qualitatively, the electrostatic potential surfaces for the fused ring systems are quite similar to those of the parent compounds shown in Figures 5.2 and 5.3. In addition, the propeller twist of perfluorotriphenylene is quite pronounced.

To calculate the binding energy of the triphenylene$\cdots$perfluorotriphenylene complex, we used the coordinates from Weck et al. [7]. Because the locations of the hydrogens were not determined in this structure, they needed to be modeled. This was done by taking the coordinates for the triphenylene ring by itself, adding the hydrogens, and optimizing their positions at the HF/6-31G** level while keeping the coordinates of the carbon atoms fixed.

We then evaluated the energetics of various aspects of the complex as outlined in Figure 5.10. We first calculated the "distortion energy" of the partners in the interaction at the HF/6-31G**, B3LYP/6-31G**, and MP2/6-31G** levels. This measures the energetic cost of twisting the triphenylene and perfluorotriphenylene rings from their ground state structures to their structures in the complex. At all three levels of theory, the distortion cost is similar, and the combined distortion

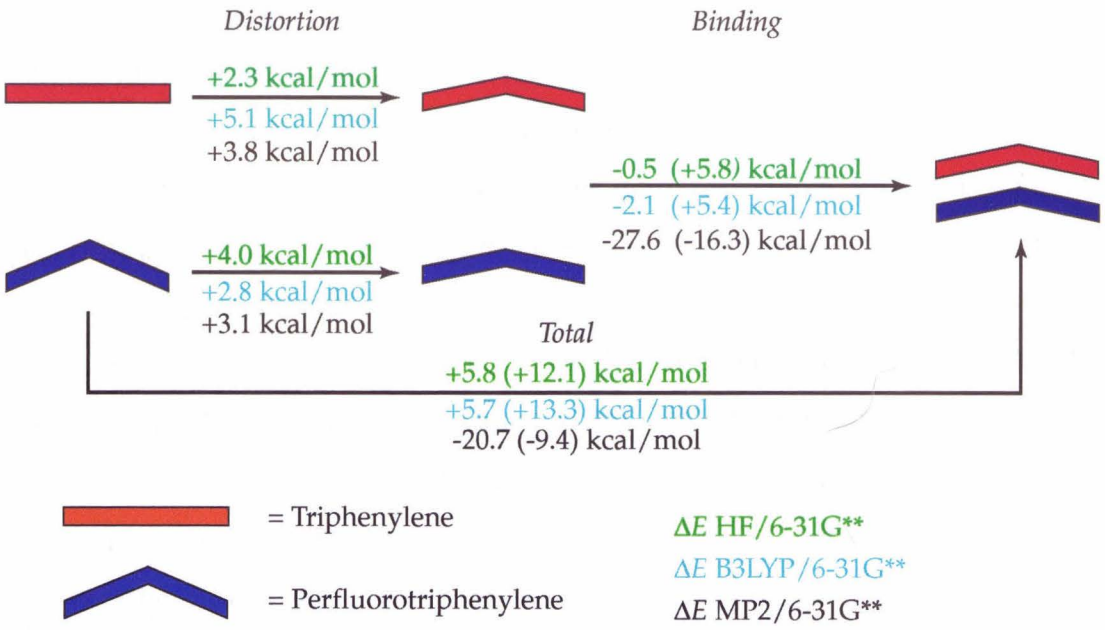


Figure 5.10: Calculated binding energies of the triphenylene$\cdots$perfluorotriphenylene complex. Numbers in parentheses are corrected for basis set superposition error. See text for a full description.

energies for both partners range from +6.3 kcal/mol to +7.9 kcal/mol, depending on the method applied.

We next calculated the binding energies of the complex at several levels of theory. This was done by calculating the total energy of the complex, and subtracting from it the total energies of the individual monomers calculated in the geometry of the complex. We first used simple supermolecule calculations and did not correct the binding energies for basis set superposition error (BSSE). The non-corrected binding energies are -0.5 kcal/mol at the HF/6-31G** level, -2.1 kcal/mol at the B3LYP/6-31G** level, and -27.6 kcal/mol at the MP2/6-31G** level. Clearly there are large differences between these levels of theory. Hartree-Fock theory and DFT both suggest very small binding energies, whereas the MP2 value is quite large. This trend is consistent with an earlier study of the benzene$\cdots$hexafluorobenzene complex [4]. When we performed counterpoise corrections to account for BSSE, the binding energies dropped significantly (these numbers are shown in parentheses in Figure 5.10). In fact, when the CP correction is included, both Hartree-Fock theory and DFT consider this interaction

to be repulsive. Finally, the total energies, including both distortion and binding are shown at the bottom of Figure 5.10. Including the effects of distortion, both Hartree-Fock theory and DFT consider the pairwise interaction to be strongly repulsive. The CP-MP2 value (-9.4 kcal/mol) seems more reasonable, especially compared to the -3.7 kcal/mol of binding energy calculated for the benzene$\cdots$hexafluorobenzene complex. It is quite surprising to see the wide range of binding energies at the different levels of theory. In the benzene$\cdots$hexafluorobenzene study, it was noted that the equilibrium distance for the partners at the MP2 level was much shorter than that at the HF and DFT levels. Perhaps, then, the distance observed in the crystal structure is more representative of the distance at the MP2 mininum than the minima at the other levels. Unfortunately, due to the size of the system, we cannot easily calculate the equilibrium binding distance at the different levels of theory. However, we can attempt analogous calculations on the benzene$\cdots$hexafluorobenzene complex to determine if similar effects are seen when using the coordinates from the crystal structure [15]. In the previous computational study, benzene and hexafluorobenzene were constrained to a C_6 symmetry axis [4]. In the crystal structure, however, the rings are not perfectly stacked, but slipped. We thus calculated the binding energies for this system as shown in Figure 5.11. In this case, the ground state structures are fully geometry optimized structures calculated at each level of theory. As evidenced by their small distortion energies, the DFT and MP2 monomer geometries are very similar to those observed in the crystal structure, whereas the Hartree-Fock monomer structures pay an energetic penalty of +6.3 kcal/mol to distort to the crystal structure geometry. In terms of binding energy, the trend mimics that seen in the triphenylene$\cdots$perfluorotriphenylene case. MP2 theory suggests a strong binding energy (-7.2 kcal/mol) for the interaction, whereas both Hartree-Fock theory and DFT report much weaker interactions (-0.2 kcal/mol and -2.1 kcal/mol, respectively). Including the effects of the counterpoise corrections weakens the binding energies further, to +2.0 kcal/mol (HF), +1.7 kcal/mol (DFT), and -3.5 kcal/mol (MP2). Interestingly, the counterpoise-corrected MP2 energy is similar in magnitude to that observed for the C_6 symmetric complex (-3.7 kcal/mol), while the geometries

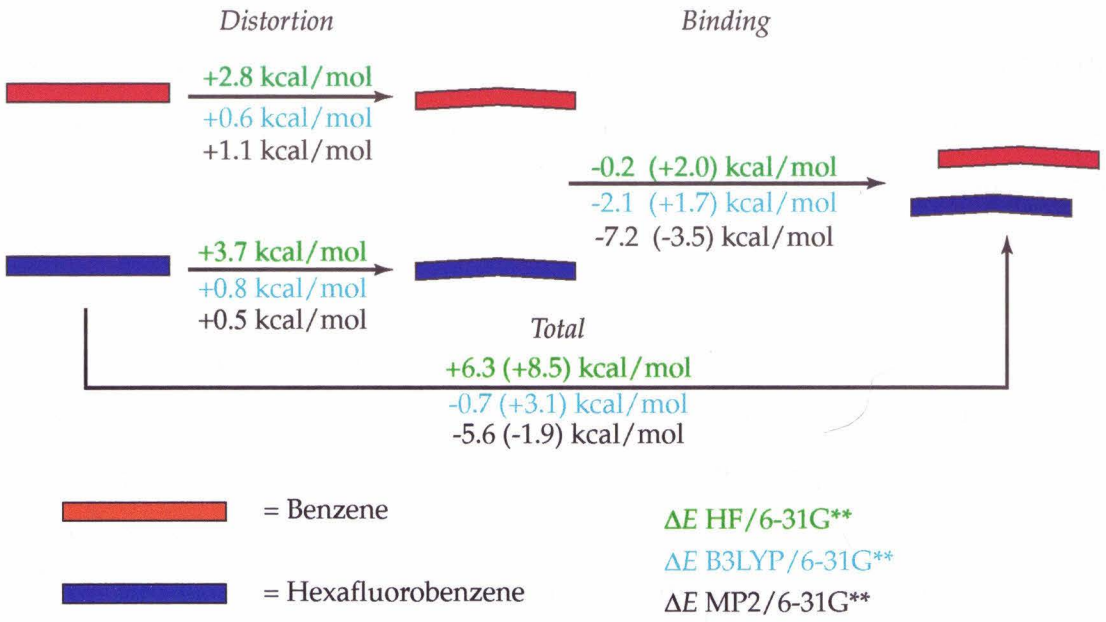


Figure 5.11: Calculated binding energies of the benzene···hexafluorobenzene complex. Numbers in parentheses are corrected for basis set superposition error. See text for a full description.

are quite different. This may suggest that the true CP-MP2 minimum corresponds to a slipped geometry, as observed in the benzene···hexafluorobenzene co-crystal. This would have implications for the use of this motif as a supramolecular synthon.[7] Further calculations may shed light on this.

Sadly, the results in this appendix probably serve as a greater testament to the weaknesses of computational methods in describing noncovalent interactions, than to our understanding of them. Nonetheless, several trends emerge from these results. The first is that Hartree-Fock theory performs poorly in describing the weak interactions between neutral molecules. This is not at all surprising given that HF theory ignores electron correlation. Perhaps somewhat more surprising is that density functional theory performs nearly as poorly as Hartree-Fock theory, despite its greater computational cost. Møller-Plesset theory seems to perform better than both HF theory and DFT. At the very least, it considers the two interactions described here as attractive. In addition, the magnitudes of the interactions seem "sensible," but only

[7]Many crystal structures of mixtures of fluorinated and non-fluorinated aryl rings exhibit the slipped-stacked geometry. See [17] for several examples.

experiment will confirm whether or not this is the case. Unfortunately, we are left with the dilemma that our "best" level of theory may not even accurately predict the results of an experiment, *and* it is prohibitively expensive computationally.[8] So where does this leave us? The major theme of this thesis is electrostatic interactions. Fortunately, electrostatics provide a predictive tool in these systems. Both of these interactions could be predicted solely by inspection of the electrostatic potential surfaces of the partners, just as in the cases of water$\cdots$benzene and water$\cdots$hexafluorobenzene. As such, although these results may reveal the weaknesses of comparatively high-level theoretical models, they undoubtedly emphasize the strengths of "electrostatic reasoning" in evaluating noncovalent interactions.

[8]The CP-MP2 binding energy for the triphenylene$\cdots$perfluorotriphenylene interaction required over 45 CPU-*days* to calculate using a Silicon Graphics 195 MHz R10000 processor.

5.6 Appendix 2 – Total Energies and Geometric Parameters

Theory	O–H bond length (Å)	H–O–H angle (degrees)	Total energy (hartrees)
HF/6-31G**	0.943	105.97	–76.0236150
MP2(full)/6-31G**	0.961	103.88	–76.2224487
B3LYP/6-31G**	0.965	103.74	–76.4197397

Table 5.3: Geometries and total energies for H_2O at various levels of theory.

Theory	C–C bond length (Å)	C–F bond length (Å)	Total energy (hartrees)
HF/6-31G**	1.378	1.314	–823.7535727
MP2(full)/6-31G**	1.391	1.340	–825.5613207
B3LYP/6-31G**	1.394	1.336	–827.5969442

Table 5.4: Geometries and total energies for C_6F_6 at various levels of theory.

Theory	Total energy
HF/6-31G**	–899.7811587
MP2(full)/6-31G**	–901.7899813
CP-MP2(full)/6-31G**	–901.7893611
B3LYP/6-31G**	–904.0205567

Table 5.5: Total energies (in hartrees) for the on-axis complex. For geometries, see Table 5.1. The energies of the monomers can be found in Tables 5.3 and 5.4. The counterpoise corrected monomer energies (in hartrees) at the CP-MP2(full)/6-31G** level are –76.2240342 and –825.5619603 for water and hexafluorobenzene, respectively.

Theory	Total energy
HF/6-31G**	–899.7818682
MP2(full)/6-31G**	–901.792246
CP-MP2(full)/6-31G**	–901.7890015
B3LYP/6-31G**	–904.023565

Table 5.6: Total energies (in hartrees) for the off-axis complex. For geometries, see Table 5.2. The energies of the monomers can be found in Tables 5.3 and 5.4. The counterpoise corrected monomer energies (in hartrees) at the CP-MP2(full)/6-31G** level are –76.22406769 and –825.5623169 for water and hexafluorobenzene, respectively.

	Triphenylene		Perfluorotriphenylene	
Theory	Ground state	Complex	Ground state	Complex
HF/6-31G**	–688.681883	–688.6780185	–1874.735386	–1874.7290964
MP2/6-31G**	–691.1236012	–691.1175665	–1879.1682995	–1879.163388
B3LYP/6-31G**	–693.2001571	–693.1921083	–1883.8631578	–1883.8586723

Table 5.7: Total energies (in hartrees) of triphenylene and perfluorotriphenylene. For triphenylene, the ground state corresponds to a calculated minimum energy structure at the given level of theory. For perfluorotriphenylene, the ground state structure is taken from the crystal structure coordinates [16] and the energy is a single-point energy calculated at the given level of theory. The structures in the complex are taken from the triphenylene$\cdots$perfluorotriphenylene co-crystal [7]. The energies are single-point energies, calculated at the given level of theory. The difference in energies between the ground state and the complex is the distortion energy reported in Figure 5.10.

		Counterpoise Corrected	
Theory	Complex	Triphenylene	Perfluorotriphenylene
HF/6-31G**	–2563.4078617	–688.6825751	–1874.7345637
MP2/6-31G**	–2570.3249285	–691.1255296	–1879.1734479
B3LYP/6-31G**	–2577.0541744	–693.1966762	–1883.8660902

Table 5.8: Total energies (in hartrees) for the triphenylene$\cdots$perfluorotriphenylene complex and for the counterpoise corrected monomers. The energies for the counterpoise corrected monomers are determined by calculating a single-point energy for the monomer in the presence of the basis functions of the other monomer, in the geometry of the complex. The differences in these values are the counterpoise corrected energies shown in parentheses in Figure 5.10.

Bibliography

[1] S. Suzuki, P.G. Green, R.E. Bumgarner, S. Dasgupta, W.A. Goddard III, and G.A. Blake. Benzene forms hydrogen bonds with water. *Science*, 257:942–945, 1992.

[2] M. Luhmer, K. Bartik, A. Dejaegere, P. Bovy, and J. Reisse. The importance of quadrupolar interactions in molecular recognition processes involving a phenyl group. *Bull. Soc. Chim. Fr.*, 131:603–606, 1994.

[3] J. H. Williams. The molecular electric quadrupole moment and solid-state architecture. *Acc. Chem. Res.*, 26:593–598, 1993.

[4] A. P. West Jr., S. Mecozzi, and D. A. Dougherty. Theoretical studies of the supramolecular synthon benzene$\cdots$hexafluorobenzene. *J. Phys. Org. Chem.*, 10:347–350, 1997.

[5] G. W. Coates, A. R. Dunn, L. M. Henling, D. A. Dougherty, and R. H. Grubbs. Phenyl-perfluorophenyl stacking interactions: A new strategy for supramolecular construction. *Angew. Chem. Int. Ed. Engl.*, 36:248–251, 1997.

[6] G.W. Coates, A.R. Dunn, L.M. Henling, J.W. Ziller, E.B. Lobkovsky, and R.H. Grubbs. Phenyl-perfluorophenyl stacking interactions: Topochemical[2+2] photodimerization and photopolymerization of olefinic compounds. *J. Am. Chem. Soc.*, 120(15):3641–3649, 1998.

[7] M. Weck, A.R. Dunn, K. Matsumoto, G.W. Coates, E.B. Lobkovsky, and R.H. Grubbs. Influence of perfluoroarene-arene interactions on the phase behavior of liquid crystalline and polymeric materials. *Angew. Chem. Int. Ed. Engl.*, 38(18):2741–2745, 1999.

[8] S. F. Boys and F. Bernardi. *Mol. Phys.*, 19:533–566, 1970.

[9] I. Alkorta, I. Rozas, and J. Elguero. An attractive interaction between the π-cloud of C_6F_6 and electron-donor atoms. *J. Org. Chem.*, 62:4687–4691, 1997.

[10] J. D. Dunitz and R. Taylor. Organic fluorine hardly ever accepts hydrogen bonds. *Chem. Eur. J.*, 3:89–98, 1997.

[11] Y. Danten, T. Tassaing, and M. Besnard. On the nature of the water-hexafluorobenzene interaction. *J. Phys. Chem. A*, 103(18):3530–3534, 1999.

[12] M. J. Frisch, G. W. Trucks, H. B. Schlegel, P. M. W. Gill, B. G. Johnson, M. A. Robb, J. R. Cheeseman, T. A. Keith, G. A. Petersson, J. A. Montgomery, K. Raghavachari, M. A. Al-Laham, V. G. Zakrzewski, J. V. Ortiz, J. B. Foresman, C. Y. Peng, P. A. Ayala, M. W. Wong, J. L. Andres, E. S. Replogle, R. Gomperts, R. L. Martin, D. J. Fox, J. S. Binkley, D. J. Defrees, J. Baker, J. P. Stewart, M. Head-Gordon, C. Gonzalez, and J. A. Pople. *Gaussian 94 (Revision D.3)*. Gaussian, Inc., 1995.

[13] A. P. Scott and L. Radom. Harmonic vibrational frequencies: An evaluation of Hartree-Fock, Møller-Plesset, Quadratic Configuration Interaction, Density Functional Theory, and Semiempirical Scale Factors. *J. Phys. Chem.*, 100:16502–16513, 1996.

[14] G. R. Desiraju. Supramolecular synthons in crystal engineering – A new organic synthesis. *Angew. Chem. Int. Ed. Engl.*, 34:2311–2327, 1995.

[15] J. H. Williams, J. K. Cockcroft, and A. N. Fitch. Structure of the lowest temperature phase of the solid benzene-hexafluorobenzene adduct. *Angew. Chem. Int. Ed. Engl.*, 31:1655–1657, 1992.

[16] A. R. Dunn and R. H. Grubbs. Unpublished results. 1999.

[17] S. A. Miller. *Magnetic Interactions in Molecular Materials.* Ph.D. Thesis, California Institute of Technology, 1999.

Chapter 6 Site-Specific Incorporation of Biotinylated Amino Acids to Identify Surface-Exposed Residues in Integral Membrane Proteins

6.1 Introduction

In this chapter, we return our attention to the nicotinic acetylcholine receptor, and outline our attempts to develop a general method to determine both the transmembrane topology of the receptor, as well as the surface accessibility of individual residues.

Determining the transmembrane topology of integral membrane proteins such as ion channels and neurotransmitter transporters is a necessary first step in evaluating their structure and function at the molecular level [1]. After topology has been established, a more demanding question is the relative exposure of individual residues to intracellular or extracellular media. Knowledge of transmembrane topology and exposure allows one to identify regions of a protein that may become glycosylated or form binding sites for external ligands, to identify intracellular regions that may be subject to phosphorylation, to assign potential binding sites for other intracellular secondary messengers such as G proteins, and to make predictions regarding the overall structure and function of an integral membrane protein.

In this chapter, we address the "higher resolution" issue of surface exposure. We wish to determine which residues are clearly on the "outside" surface of a receptor. Of course, if a residue is determined to be exposed, the topology question is answered. Residue accessibility is clearly a related measure of protein structure, but, as discussed below, addresses a somewhat different issue. We consider here a relatively well understood membrane protein, the nicotinic acetylcholine receptor.[1] There is general agreement concerning the overall topology of this receptor, although final proof will require high resolution structural data. Each subunit is thought to have a large, extracellular N-terminal domain, followed by four transmembrane regions and a necessarily extracellular C-terminus (Figure 6.1A). However, little is known about the relative exposures of residues throughout the receptor.

Many other membrane channels, transporters, and receptors have less well understood topologies. The first step (and, unfortunately, often the last step) in determining the topology is the generation of a hydropathy plot based on the protein

[1]For an introduction to the nicotinic acetylcholine receptor, please see Chapter 2.

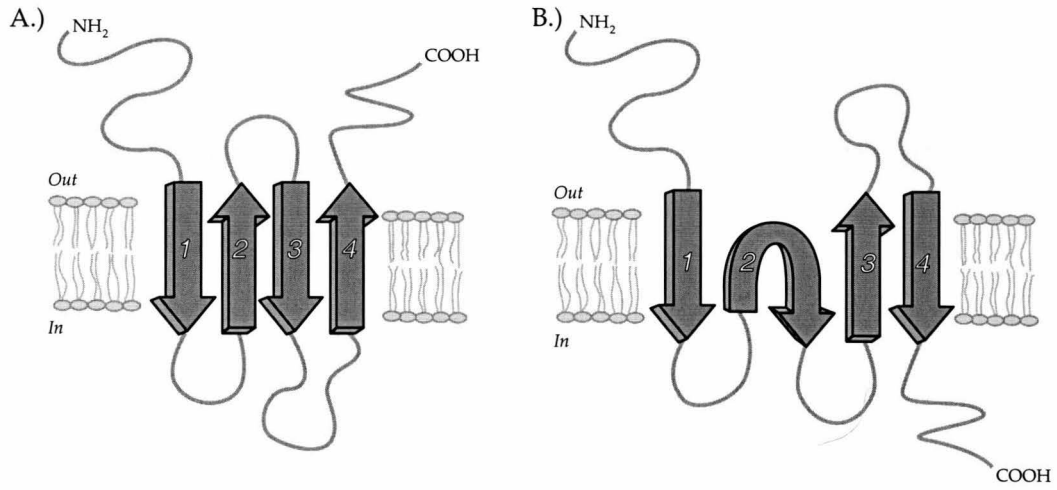


Figure 6.1: A.) Generally accepted topology of each subunit of the nicotinic acetylcholine receptor. This is also the original topology model (1989–1995) proposed for the glutamate receptor, GluR3 [2]. B.) The revised topology for the glutamate receptor [3, 4].

sequence. Although hydropathy plots are useful in suggesting a possible topology for a membrane protein, they are by no means an effective substitute for experimental evidence. For instance, the primary amino acid sequence for the ionotropic glutamate receptor was first determined in 1989 and led to a topological model analogous to that for the nicotinic receptor – four transmembrane domains and both the N and C termini located extracellular [2]. Subsequent work showed that the putative second transmembrane domain most likely forms a reentrant loop, so that the N and C termini are on the opposite sides of the membrane (Figure 6.1B) [3, 4]. Such a drastic change certainly alters ones view of the structure and mode of action of the receptor. Other recent experiments on a variety of ion channels and transporters, including the potassium channel, ROMK1 [5]; the GABA transporter, GAT-1 [6]; the glycine transporter, GLYT1 [7]; and the Na^+/glucose cotransporter, SGLT1 [8] have suggested inconsistencies in the transmembrane topologies predicted from hydropathy analysis.

These studies underscore the importance of having definitive experimental approaches to determine transmembrane topology of membrane proteins. Although several methods are available, a new approach would be welcome if it overcomes

some of the inherent weaknesses of the existing methods.

6.2 Previous Approaches to Determining Topology and Accessibility

A wide variety of techniques have been developed to address the topology issue [1]. We will briefly describe below some of the higher precision approaches that target specific regions of integral membrane proteins.

6.2.1 Endogenous Glycosylation Sites and N-Glycosylation Scanning Mutagenesis.

One of the most common methods to determine transmembrane topology is to assay a particular region of a protein for endogenous N-linked glycosylation. Because N-linked glycosylation occurs, with few exceptions [9], on the luminal side of the endoplasmic reticulum, only regions destined to become extracellular are glycosylated. Thus, identifying a site that is glycosylated in the wild-type protein is good evidence that it is located extracellularly. Expanding upon this tactic, one can engineer N-linked glycosylation sites into transmembrane proteins to determine topology. This approach has been used recently to assay the transmembrane topologies of many ion channels and transporters, including the GABA transporter, GAT-1 [6]; the glycine transporter, GLYT1 [7]; CFTR [10]; and the Na^+/glucose cotransporter, SGLT1 [8]. In all of these studies, the functional competence of the glycosylated mutants was often quite poor compared with that of the wild-type protein, and in many cases the modified proteins did not function at all. This is of particular concern because the introduction of a non-native glycosylation site may alter topology [4]. If a particular protein is found to be glycosylated in the experiment but does not function, it is often difficult to determine whether the protein fails to function because glycosylation interferes with the normal function of the protein, or because glycosylation has caused the protein to fold with an altered topology.

6.2.2 Cysteine Accessibility/SCAM

Another common method for determining transmembrane topology is the introduction of cysteine residues into the protein of interest, followed by labeling with thiol-reactive reagents. If the reagent is membrane impermeant, topology can be evaluated by comparing the results of intracellular vs. extracellular application [11, 12]. For example, cysteines have been modified with biotin reagents, and the reaction detected by streptavidin binding [13, 14]. The cysteine modification approach has an advantage over glycosylation methods in that a single cysteine mutant is often less perturbing to the overall protein structure than the introduction of a glycosylation site. The method has been generalized by Karlin to SCAM (substituted-cysteine accessibility method), allowing the "accessibility" of specific residues to be probed with relatively smaller thiol-reactive reagents [15, 16]. Note that such accessibility is quite different from "exposure" as used here. SCAM has labeled residues that nominally lie in a transmembrane region, but that are nonetheless accessible because they lie along an ion channel or within an agonist binding site. It is in this regard that we distinguish accessibility vs. surface exposure. Potential disadvantages of these methods arise from the contributions of endogenous cysteine residues that may also be labeled in the assay. In addition, the introduction of new cysteines may lead to improper disulfide formation and incorrectly folded proteins. A possible solution to these problems is to mutate all of the endogenous cysteines to another amino acid, and to then introduce a unique cysteine. This has been successful in some proteins, but there are certainly cases where removal of all the endogenous cysteines will render the protein nonfunctional. Furthermore, for an in vivo assay, there may be complications arising from reactive cysteines found in endogenous cellular proteins that cannot be mutated away.

6.2.3 Epitope Protection

In this assay, a fusion is created between the protein of interest and a well defined epitope [4, 17, 18]. The proteins are synthesized in vitro in the presence of microsomal

vesicles so that the expressed protein is inserted into the membrane. Digestion of the translation mixture with proteases cleaves extramicrosomal (cytoplasmic) regions, destroying the epitope if it is displayed in this region. If the epitope is located inside the vesicle, it is protected from protease digestion and can be immunoprecipitated or detected in a Western blot. There are several concerns raised by this assay. The assay is performed in a cell-free system; therefore the topology may not be identical to that produced in an intact cell [18]. The second concern is that the epitope introduced is often quite large (the prolactin epitope is 195 amino acids!) and consequently must often be introduced as a C-terminal fusion to a truncated and often non-functional protein [4]. Finally, there is always a concern that the introduction of a large epitope may cause the protein of interest to fold improperly. This potential problem, coupled with the fact that the proteins often cannot be assayed for function, makes the assay less attractive.

6.2.4 Tyrosine Iodination

Hucho and co-workers have used [^{125}I]-labeling to probe surface exposure of tyrosine residues of the nicotinic receptor harvested from *Torpedo californica* [19]. This method has some attractive features, but, since all exposed tyrosines are labeled simultaneously, one must sequence the labeled protein to determine exposure. This is not feasible with proteins for which there is not a naturally occurring plentiful source.

6.3 The Biocytin Approach

To address the potential shortcomings of the present methods, we sought to develop a new experimental approach to discover surface exposed residues. Necessarily, such information will bear strongly on issues of transmembrane topology and accessibility. Our design goals were as follows:

- To determine the surface exposure/topology of a protein in vivo
- To assay functional proteins with wild-type behavior

- To introduce a single, minimally perturbing mutation
- To use a marker that is endogenous neither to the protein of interest nor to other cell surface proteins
- To detect the marker using robust and readily available methods
- To produce a graded series of reagents that allows conclusions about relative exposures

As a first step, we have combined two commonly used techniques: the biotin-streptavidin interaction, and the in vivo nonsense suppression method.

Biotin is a small molecule that binds with high affinity ($K_d \approx 10^{-15}$ M) to streptavidin, a 60 kD tetrameric protein. Biotin can be readily functionalized at its carboxy-terminus while retaining its ability to bind tightly to streptavidin, rendering it an ideal probe [20–22]. Biotinylated amino acids have previously been incorporated into proteins in vitro by supplementing the translation mixture with full-length lysyl-tRNA functionalized with biotin derivatives at the ε-amino group of lysine, thus leading to random incorporation of biotinylated amino acids at various lysine sites in the protein [23]. Although this suggested that biosynthetic incorporation of biotinylated amino acids was possible, the remaining challenge was to introduce the biotin label site-specifically in the protein of interest in a living cell. To accomplish this we employed the nonsense suppression method for incorporation of unnatural amino acids into proteins [24–27] as adapted in our labs to the in vitro heterologous expression system of the *Xenopus* oocyte [28–30].

In the in vivo nonsense suppression method, the oocyte is co-injected with 2 mutated RNA species: 1) mRNA synthesized in vitro from a mutant cDNA clone containing a stop codon, TAG, at the amino acid position of interest; and 2) a suppresser tRNA containing (a) the complementary anti-codon sequence (CUA) and (b) the desired unnatural amino acid chemically acylated to the 3′ end. During translation by the oocyte's endogenous synthetic machinery, the unnatural amino acid is specifically incorporated at the appropriate position in the protein encoded by the

mRNA. Recent studies from these labs have shown that this method is broadly applicable, being compatible with many different proteins, a large number of sites in a given protein, and a wide range of amino acids [28, 30].

As a test amino acid we chose biocytin, a naturally occurring (but not translationally incorporated) biotinylated amino acid. Biocytin's relatively small size, similarity to the natural amino acid lysine, and commercial availability made it particularly attractive.

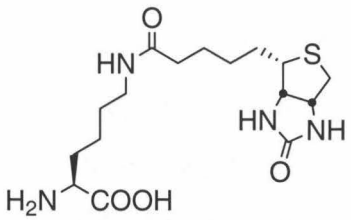


Figure 6.2: Structure of biocytin

To test the method, we chose the prototype ligand-gated ion channel, the nicotinic acetylcholine receptor (nAChR), a 290 kD pentameric ion channel found at the neuromuscular junction [31–35]. In particular, we chose to evaluate the main immunogenic region (MIR) of the α subunit. The MIR is composed of ten residues in the N-terminal region of the α subunit (residues 67-76) and binds many antibodies, including those associated with myasthenia gravis, a human autoimmune disease [36–38]. Furthermore, this region has been shown by electron microscopy to be located near the synaptic end of the nAChR [39]. Because it is known to be topologically extracellular and accessible as a unit to large proteins, the MIR appeared to be an ideal location to test the biocytin methodology for evaluating the surface exposure of individual residues. Given the tight fit between biotin and streptavidin, we anticipated that only the most highly exposed positions of the MIR would form strong biocytin/streptavidin complexes.

6.4 Results and Discussion

We first mutated the codons for each of the 10 residues of the main immunogenic region of the nAChR from mouse muscle (α67-76 – see Table 6.1 for sequence) to the stop codon TAG. We then co-injected *Xenopus* oocytes with a mixture of mutant mRNA (1.2 ng total/oocyte, 4:1:1:1 α:β:γ:δ subunit stoichiometry) containing the TAG codon at a single position in the α subunit, and an aminoacyl-tRNA (25 ng/oocyte) charged with the amino acid of interest. After 36-48 hours, oocytes were assayed electrophysiologically for response to 200 μM acetylcholine (ACh) by the two electrode voltage clamp technique [40]. The results from a successful biocytin incorporation are shown in Figure 6.3. In this experiment, biocytin incorporation was successful at both positions α70 and α76. As controls, suppression with a natural amino acid, tyrosine at α76, and experiments using uncharged tRNA (tRNA-dCA) [28–30] were performed as well. Clearly, the biotinylated mutants functioned at levels comparable to those for mutants suppressed with a natural amino acid. In addition, the very small currents arising from the dCA control (uncharged tRNA) demonstrate that the currents observed in the biocytin experiments arise solely from receptors containing the biotinylated amino acid.

At five of the ten positions in the MIR – α68, 70, 71, 75 and 76 – biocytin was incorporated efficiently into functional receptors with whole-cell currents ranging from several hundred to several thousand nA. At the other five positions – α67, 69, 72, 73 and 74 – whole-cell currents were generally less than 200 nA. In each experiment, the α70 biocytin mutant was expressed as a positive control so that whole-cell currents, which often vary substantially between oocyte batches, could be compared to an internal standard. A summary of these experiments is shown in Figure 6.4.

Again as a control, uncharged tRNA was tested at all the sites. In all cases, the current evoked from 200 μM ACh was less than 20% of that obtained in the suppression experiment, and in nine out of ten cases, it was less than 5%.

The electrophysiology experiments clearly established the presence of functional, biocytin-containing ACh receptors on the oocyte surface for 5 of the 10 positions in

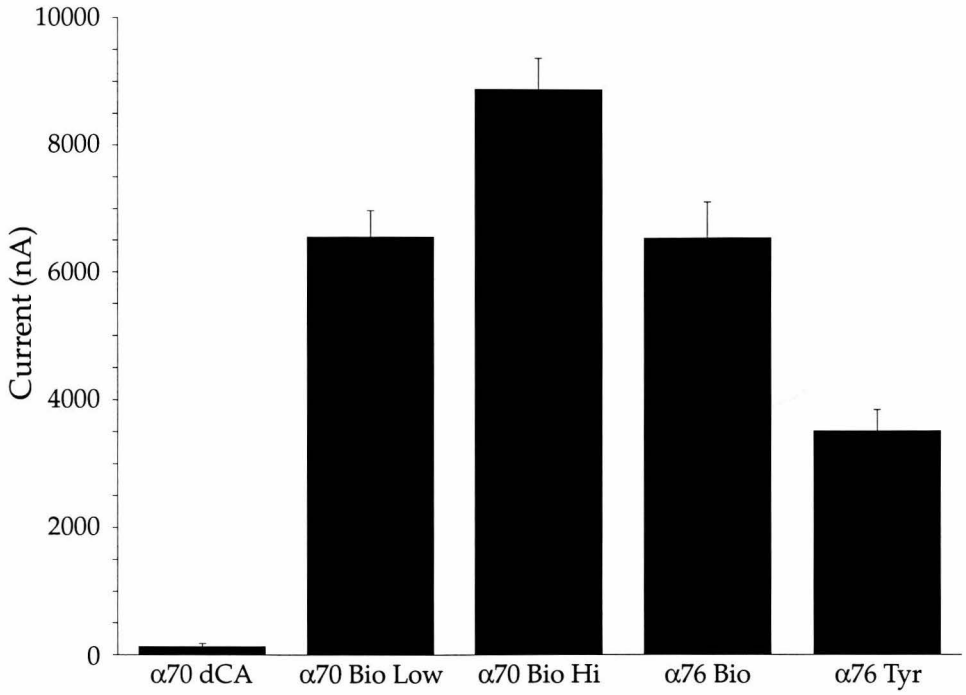


Figure 6.3: Average whole-cell currents in response to 200 μM ACh for oocytes injected with nAChR α70TAG or α76TAG (4:1:1:1, 1.2 ng) and uncharged tRNA (dCA), or tRNA charged with biocytin (Bio) or tyrosine (Tyr). The two bars for the α70 biocytin are the averages of the currents for the five lowest expressing oocytes and the five highest expressing oocytes. Error bars are $\pm$ s.e.m. (n=5)

the MIR. Sequence conservation analysis of 34 nAChR α subunits shows that biocytin produced functional receptors at five of the six *least* conserved positions in the MIR (Table 6.1). The fact that biocytin did not express well at the sixth, αGly73, may arise because a glycine (side chain molecular weight=1) to biocytin (side chain molecular weight=298) mutant may epitomize the phrase "non-conservative mutation." Furthermore, the four most highly conserved residues, three of which are greater than 70% conserved across *all* nicotinic subunits, are the positions at which biocytin did not express. Since these four residues are highly conserved, they may play crucial structural or functional roles in the receptor, and even non-conservative conventional mutations may not be tolerated. We believe that the low currents observed from attempted biocytin incorporation at these positions indicate that the relatively large side chain of biocytin is not tolerated at these positions.

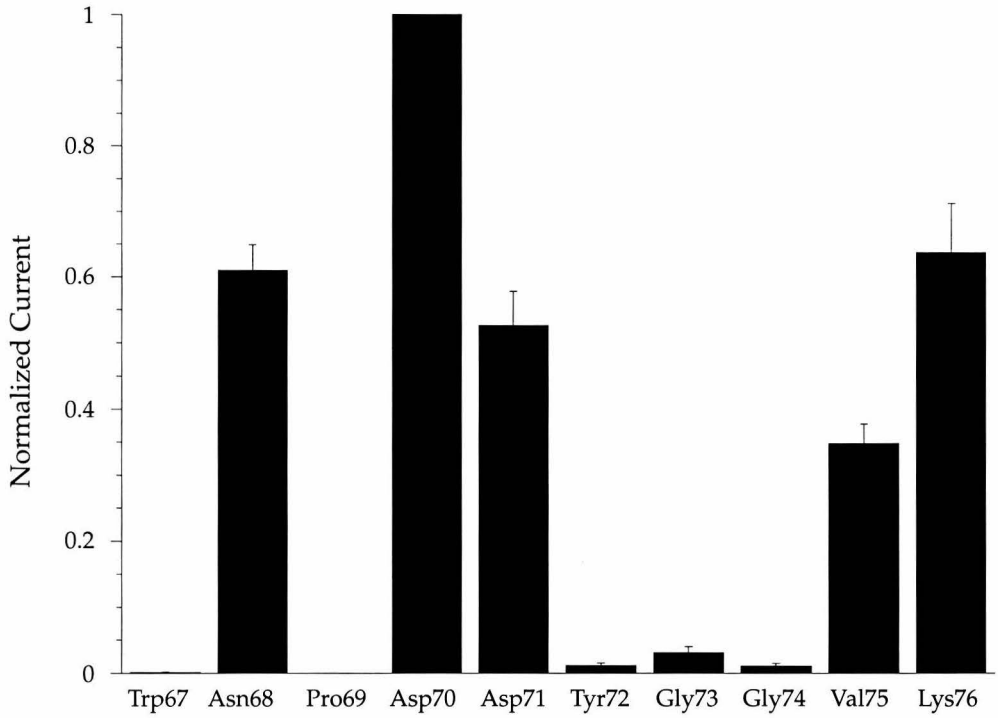


Figure 6.4: Whole-cell currents in response to 200 μM ACh from attempted biocytin suppression experiments at each of the 10 positions of the MIR of the nAChR. Results are normalized to the whole-cell current of α70 biocytin expressed in the same experiment. Bars represent the mean of two experiments, each with five oocytes.

With electrophysiological evidence for functional biotinylated receptors, we turned our attention to binding streptavidin to assay the surface exposure of particular side chains. To assay for streptavidin binding, batches of five oocytes whose whole-cell currents had previously been determined were incubated in a solution of 300 pM [^{125}I]-streptavidin for two hours, followed by extensive washing (see Experimental section). Individual oocytes were then measured in a γ-counter. Of the five mutants that showed large whole-cell currents, only one, α70 biocytin, showed streptavidin binding above background levels. In addition, no streptavidin binding was observed with any of the five mutants that did not give large whole-cell currents, or to cells expressing non-biotinylated nAChRs. The results of a typical experiment are shown in Figure 6.5.

In the case of α70 biocytin, the mutant that showed streptavidin binding, we also evaluated the binding of [^{125}I]-α-bungarotoxin – a snake toxin that is highly

Position	Percent conservation in α subunits	ACh response with biocytin incorporation
Trp67*	97	—
Asn68	56	+
Pro69	79	—
Asp70	26	+
Asp71	53	+
Tyr72*	91	—
Gly73	47	—
Gly74*	82	—
Val75	62	+
Lys76	44	+

Table 6.1: Sequence analysis of the MIR from 34 nAChR α subunits in the SWISS-PROT database. Positions marked with a * are $> 70\%$ conserved across all nicotinic subunits (i.e., not solely α subunits).

specific for the nAChR. We found that the amounts of streptavidin and bungarotoxin binding were similar, suggesting that all nAChRs were completely labeled by both bungarotoxin and streptavidin. This also provides concrete biochemical evidence to support the electrophysiological evidence that the in vivo suppression method produces full length protein incorporating only the desired amino acid. If this were not the case, i.e., if some nAChR was produced with a natural amino acid incorporated at the suppression site due to read through of the stop codon or reacylation of the tRNA, we would have expected the amount of bungarotoxin binding to exceed the amount of streptavidin binding.

To explore whether the streptavidin binding sites and bungarotoxin binding sites overlap, we incubated oocytes expressing nAChR α70 biocytin with unlabeled streptavidin. We then applied [^{125}I]-α-bungarotoxin to these oocytes to determine if it could still bind. We also performed the reverse experiment, incubating oocytes expressing nAChR α70 biocytin with unlabeled α-bungarotoxin and attempting to bind [^{125}I]-streptavidin. The results of these experiments are shown in Figure 6.6.

These results clearly demonstrate that binding of bungarotoxin does not strongly inhibit binding of streptavidin to the α70 biocytin mutant and vice versa. This is consistent with previous studies that demonstrated that both bungarotoxin and

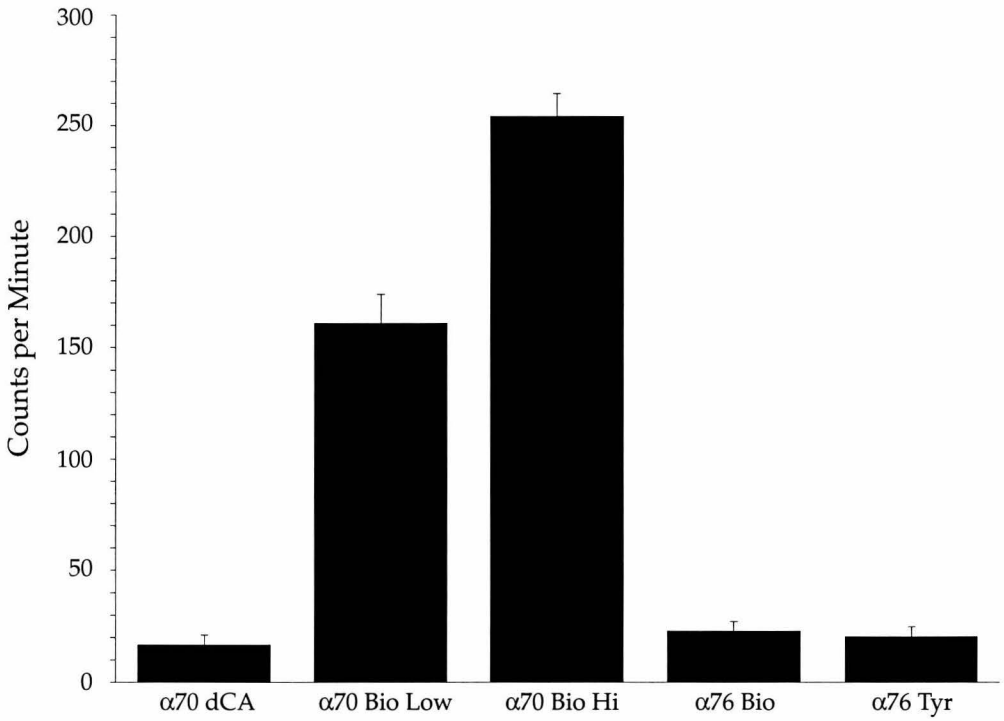


Figure 6.5: [^{125}I]-Streptavidin binding to oocytes injected with nAChR α70TAG or α76TAG (4:1:1:1, 1.2 ng) and uncharged tRNA (dCA), or tRNA charged with biocytin (Bio) or tyrosine (Tyr). The oocytes used in this experiment are identical to those whose whole-cell currents are shown in (Figure 6.3), thus demonstrating that binding of streptavidin scales with whole-cell current. The α76 Tyr mutation establishes that streptavidin does not bind to non-biotinylated nAChRs. Detector background has been subtracted in all cases. Error bars are $\pm$ s.e.m. (n=5)

antibodies against the MIR are capable of binding simultaneously to the intact receptor [41].

Finally, we have measured the EC_{50} value for ACh for the α70 biocytin mutant both before binding streptavidin ($EC_{50} \approx 39$ μM) and after binding ($EC_{50} \approx 34$ μM). These values are not significantly different from one another, nor are they significantly different from EC_{50} of the wild-type receptor (49 μM). At first it may seem surprising that binding a large protein to the receptor would not alter its pharmacological behavior. However, previous studies have shown that antibodies binding to the MIR do not substantially alter the physiology of the receptor [42].

To gain a clearer physical picture of the above results, we have considered them in light of a previously published structural model. Tsikaris et al. have used NMR

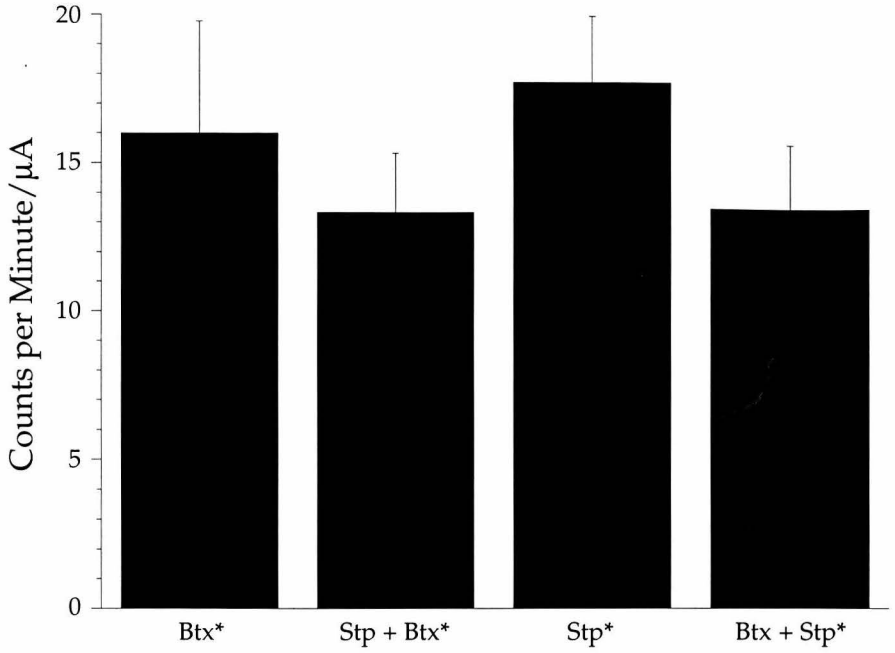


Figure 6.6: Whole-cell binding of [^{125}I]-streptavidin and [^{125}I]-bungarotoxin to AChR α70 biocytin mutants. Btx*: binding [^{125}I]-bungarotoxin to untreated oocytes; Stp + Btx*: binding [^{125}I]-bungarotoxin to oocytes pretreated with streptavidin; Stp*: binding [^{125}I]-streptavidin to untreated oocytes; Btx + Stp*: binding [^{125}I]-streptavidin to oocytes pretreated with bungarotoxin. Within the error bars shown, the data demonstrate that [^{125}I]-bungarotoxin and [^{125}I]-streptavidin do not interfere with one another in binding to the α70 biocytin nAChR. Binding is expressed as counts per minute per μA of whole-cell current to account for different expression levels between groups of oocytes. Counts have been normalized to account for the differences in specific activity between the [^{125}I]-streptavidin and [^{125}I]-bungarotoxin. Error bars are $\pm$ s.e.m. (n=5)

to determine the structure of a ten-residue peptide corresponding to the sequence of the MIR from the α subunit of the nAChR from *Torpedo californica* [43]. This sequence differs from that of the mouse muscle MIR studied here at two sites – Ala70 and Ile 75. Starting with the coordinates for the *Torpedo* MIR taken from the Protein Data Bank [44], we altered the side chain identities of these two residues to the corresponding mouse muscle residues (Asp and Val, respectively) using standard geometries. This structure is shown in Figure 6.7. We have color-coded the structure as follows: purple – sites at which biocytin incorporation was unsuccessful; gray – sites that incorporated biocytin but did not bind streptavidin; red – α70, the sole

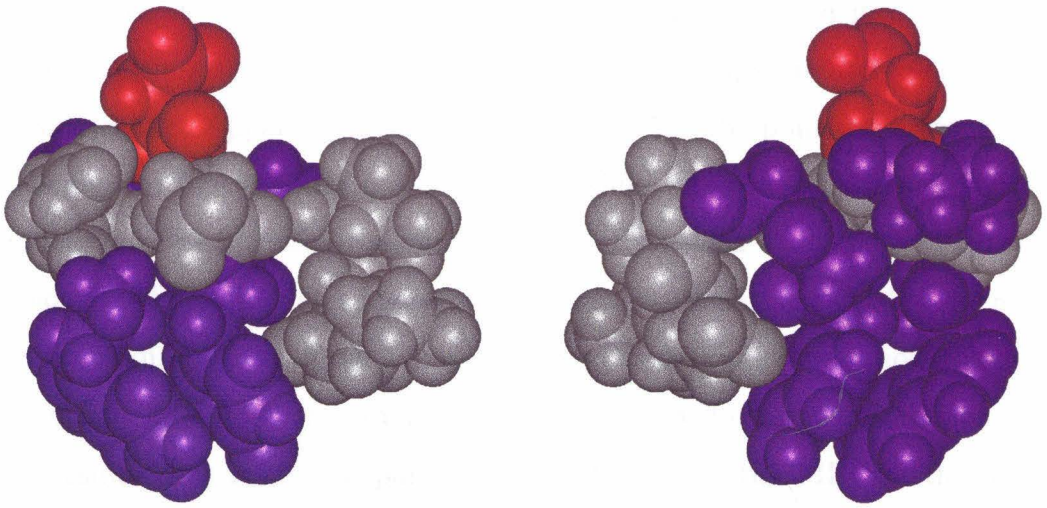


Figure 6.7: NMR structure (CPK model) of MIR peptide (PDB:1TOR) with modifications described in text. Color scheme: purple – sites at which biocytin incorporation was unsuccessful; gray – sites that incorporated biocytin but did not bind streptavidin; red – α70, the sole streptavidin-binding site. The right view is derived from the left by a 180° rotation around the vertical (y) axis.

streptavidin-binding site.

The structure in Figure 6.7 leads naturally to a rationalization of our results. The MIR peptide clearly has two sides – one that can tolerate biocytin, and one that cannot. Biocytin is larger than any of the 20 natural amino acids. Perhaps it will produce functional receptors only if there is a pre-existing (or easily created) open space for this larger side chain. We suggest that the "gray" side has this open space, and so presumably faces out away from the bulk of the protein. In contrast, the "purple" side cannot tolerate biocytin, and so apparently these side chains project down into the rest of the protein.

The selectivity of the biocytin method is shown by the result that biocytin was incorporated at five sites, yet only one of these sites was competent to bind streptavidin under the conditions of these experiments. The crystal structure of the biotin-streptavidin complex reveals that the ligand is buried fairly deeply in the protein, with even the terminal carboxylate not fully exposed [45]. Only residues that position the biocytin in an unencumbered environment will lead to streptavidin binding. As shown in Figure 6.7, the Asp70 side chain (red) points "up", away from, for example, the

Trp67 or Try72 side chains that are postulated to project down into the receptor. The other "biotin-accepting" side chains (gray) project approximately normal to the Asp70 side chain. Apparently, these sites (68, 71, 75, 76) point to a space that is large enough for biocytin but not for streptavidin. There is precedent for binding interactions with dissociation constants significantly weaker than 10^{-15} M for substituted biotins [46], and perhaps higher streptavidin concentrations would have allowed detectable binding at the "gray" sites. However, such studies were not feasible due to non-specific binding of streptavidin to the oocytes at higher concentrations.

Our results thus suggest a particular orientation for the MIR on the nAChR. Residue α70 points directly out away from the receptor. The side chains emanating from residues 67, 69, and 72 (73 and 74 are Gly) project down into the protein. The remaining residues, 68, 71, 75, and 76, project into open space, but perhaps toward some other structural feature of the receptor that prevents streptavidin binding.

This analysis implies that a tether longer than is present in biocytin might allow streptavidin binding at the "gray" sites. We therefore prepared two longer-chain analogs of biocytin, A and B. Unfortunately, we were not able to incorporate the longer analog B, into functional receptors. We could incorporate biocytin-analog A at position α70 at levels that were readily detectable by electrophysiology (500-1000 nA). This is a remarkably "unnatural" amino acid, suggesting considerable flexibility in the ribosomal machinery. However, this amount of expression was not sufficient for the less sensitive streptavidin assay, and so we were not able to make use of the longer chain analog.

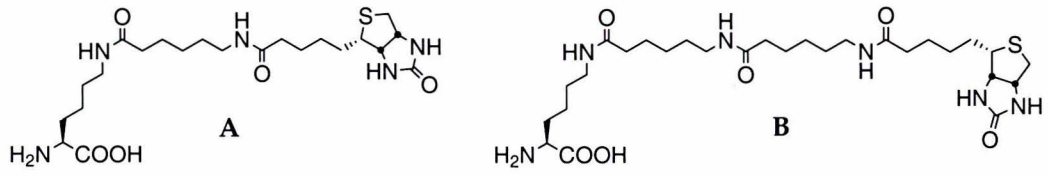


Figure 6.8: Biocytin analogs

We conclude that the method of site-specific biocytin incorporation with streptavidin detection provides a potentially general way to identify highly surface exposed residues in complex, integral membrane proteins. It thus provides information that

goes beyond the "simpler" question of transmembrane topology. This technique will likely be applicable to other channels, receptors, and transporters. Our data also show that biocytin derivatives with longer tethers do not express well, so that the biocytin method alone cannot be simply generalized to a wide range of exposures. Alternative strategies can be envisioned, perhaps involving incorporation of unnatural residues that are rather small but that have specific reactivities allowing attachment of biotin moieties with longer chains [47–49].

6.5 Summary

The in vivo nonsense suppression method for the incorporation of unnatural amino acids is a powerful tool for obtaining chemical-scale information about functional proteins expressed in *Xenopus* oocytes. We have extended the method to incorporate a biotinylated amino acid site-specifically within the main immunogenic region (MIR) of the nicotinic acetylcholine receptor (nAChR), the prototypical ligand-gated ion channel. Using [^{125}I]-labeled streptavidin as a probe, we have identified an amino acid in the MIR that is highly exposed on the surface of the protein. We believe that this method represents an improvement over present approaches to determining surface exposure and transmembrane topology for a number of reasons. In particular, the method does not require purified protein; uses a minimally perturbing point mutation; uses a marker that is not endogenous to either the protein of interest or to other cell surface proteins; allows detection by a commercially available assay; and determines the topology/surface exposure in a functional, characterizable protein expressed in vivo. For these reasons, we consider the method to be a potentially general one for determining the surface exposure of specific residues of transmembrane proteins expressed in vivo. In a broader sense, these results further demonstrate the power of the in vivo nonsense suppression method to apply the tools of organic chemistry to the problems of molecular neuroscience.

6.6 Experimental

Reagents were purchased from Aldrich unless otherwise noted. (3-[^{125}I]-iodotyrosyl–54)-α-bungarotoxin (2000 mCi/mmol) and [^{125}I]-streptavidin (30 mCi/mg) were purchased from Amersham. Anhydrous THF was distilled from Na/benzophenone; anhydrous DMF was obtained from Aldrich. All chemical syntheses were performed under a positive pressure of argon unless noted otherwise. NMR spectra were recorded on a General Electric QE300 (300 MHz ^{1}H, 75 MHz ^{13}C), JEOL Delta 400 (100 MHz, ^{13}C), or a Bruker AM-500 (500 MHz ^{1}H, 125 MHz ^{13}C) spectrometer and are referenced to residual solvent protons. Mass spectrometry (FAB) was performed by the mass spectrometry facilities at the University of California, Riverside, or at Caltech. UV/Vis spectra were recorded on a Beckman Instruments DU-640 continuous-wave spectrometer. HPLC separations were performed on a Waters dual 510 pump liquid chromatograph system equipped with either a Waters 490E variable wavelength UV detector or a Waters 960 photo diode array detector. Analytical HPLC samples were separated using a Waters NovaPak C_{18} column (3.9 x 150 mm); semi-preparative samples were separated using a Whatman Magnum 9 column (9.4 x 500 mm, Partisil 10, ODS-3). Oligonucleotides for mutagenesis were synthesized by the Biopolymer Synthesis Facility at Caltech. PCR reactions were performed using a Perkin-Elmer TC-1 DNA Thermal Cycler. Automated sequencing of PCR products was performed by the DNA Sequencing Core Facility at Caltech. Ligation of aminoacyl-dCA to tRNA-THG73(-CA) was performed as described previously [29, 30].

6.6.1 Synthesis of Biotinylated Amino Acids

NVOC-Biocytin Cyanomethyl Ester

Biocytin (50 mg, 0.13 mmol, Molecular Probes) and sodium carbonate (28 mg, 0.27 mmol) were dissolved with stirring in water (7.5 mL) and freshly distilled THF (5 mL). A solution of 6-nitroveratryloxycarbonyl chloride (NVOC-Cl) (37 mg, 0.13 mmol) in 5 mL THF was added dropwise [50, 51]. After three hours, solvents were removed in

vacuo and the remaining residue was dissolved in dry DMF (1.5 mL) and chloroacetonitrile (1.5 mL). Triethylamine (400 μL) was added and the mixture was stirred overnight. Volatiles were removed in vacuo and the remaining solid was purified by flash chromatography (silica gel, CH_2Cl_2 + MeOH 1-10%) to give 67 mg of NVOC-biocytin cyanomethyl ester as a pale yellow solid (77% yield for 2 steps): ^{1}H NMR (500 MHz, DMSO-d_6) δ 8.10 (d - br, J=6.8 Hz, 1H), 7.74 (t - br, J=5.0 Hz, 1H), 7.70 (s, 1H), 7.18 (s, 1H), 6.39 (s, 1H), 6.33 (s, 1H), 5.38 (AB, J =15.0 Hz, 1H), 5.34 (AB, J =15.0 Hz, 1H), 5.01 (s, 2H), 4.27 (m, 1H), 4.11 (m, 1H), 3.92 (s, 3H), 3.86 (s, 3H), 3.13 (m, 1H), 3.04 (m, 1H), 2.99 (m, 2H), 2.80 (dd, J =10 Hz, 5.5 Hz, 1H), 2.57-2.46 (m, 3H), 2.03 (t, 2H), 1.70-1.15 (m, 10H); ^{13}C NMR (75 MHz, DMSO-d_6 δ 172.0, 171.6, 162.8, 155.8, 153.5, 147.8, 139.2, 127.8, 115.8, 110.3, 108.2, 62.8, 61.1, 59.3, 56.3, 56.1, 55.5, 53.6, 49.5, 39.9, 38.1, 35.3, 30.2, 28.7, 28.3, 28.1, 25.4, 22.8.

α-NVOC-ε-(6-(biotinoyl)amino)hexanoyl-L-lysine (NVOC-A)

A solution of α-NVOC-lysine (42 mg, 0.11 mmol, 1.0 eq., prepared in two steps from α-Boc-lysine and NVOC-Cl, followed by removal of the Boc-group with TFA), 6-((biotinoyl)amino)caproic acid N-hydroxysuccinimide ester (50 mg, 0.11 mmol, 1.0 eq., Molecular Probes) and Na_2CO_3 (47 mg, 0.44 mmol, 4.0 eq.) was stirred in dry DMF (5 mL) under argon. After 3 days, the mixture had become gel-like and a solution of 1N $NaHSO_4$ (50 mL) was added, followed by CH_2Cl_2 (50 mL). The mixture was stirred and a pale yellow precipitate formed. The solid was filtered, washed with CH_2Cl_2, water, acetone, and dried to afford 42 mg of a pale yellow solid (53% yield): ^{1}H NMR (500 MHz, DMSO-d_6) δ 7.74 (d - br, J =8 Hz, 1H), 7.66 (m, 2H), 7.64 (s, 1H), 7.12 (s, 1H), 6.39 (s, 1H), 6.33 (s, 1H), 5.33 (AB, J =15.1Hz, 1H), 5.25 (AB, J =15.1Hz, 1H), 4.23 (t, J =5.5 Hz, 1H), 4.06 (m, 1H), 3.85 (s, 3H), 3.80 (s, 3H), 3.02 (m, 1H), 2.93 (m, 4 H), 2.82 (s, 1H), 2.76 (m, 1H), 2.73 (s, 1H), 2.50 (d, J =12.4 Hz, 1H), 1.94 (m, 4H), 1.63 (m, 1H), 1.52 (m, 2H), 1.37 (m, 4H), 1.28 (m, 8H), 1.15 (m, 2H); ^{13}C NMR (100 MHz, DMSO-d_6) δ 173.8, 171.9, 171.8, 162.7, 155.8, 153.5, 147.6, 139.0, 138.1, 128.2, 110.0, 108.1, 62.4, 61.0, 59.2, 56.2, 56.1, 55.4, 53.8, 38.3, 38.1, 35.4, 35.2, 30.4, 28.9, 28.7, 28.2, 28.0, 26.1, 25.3, 25.0, 23.0.

α-NVOC-ε-(6-(biotinoyl)amino)hexanoyl-L-lysine Cyanomethyl Ester

To a solution of α-NVOC-ε-(6-(biotinoyl)amino)hexanoyl-L-lysine (40 mg, 0.055 mmol, 1.0 eq.) in 2 mL dry DMF was added triethylamine (24 μL, 0.165 mmol, 3.0 eq.) and chloroacetonitrile (2 mL). After stirring for 9 hours under argon, volatiles were removed in vacuo and the remaining solid was partitioned between water and CH_2Cl_2. A pale yellow solid remained and was combined with the organic fraction, concentrated and dried in vacuo to afford the cyanomethyl ester as a pale yellow solid (20 mg, 47%): ^{1}H NMR (500 MHz, DMSO-d_6) δ 8.07 (d, J =7.5 Hz, 1H), 7.70 (m, 2H), 7.18 (s, 1H), 6.39 (s, 1H), 6.33 (s, 1H), 5.38 (AB, J =14.8 Hz, 1H), 5.34 (AB, J =14.8 Hz, 1H), 5.01 (s, 2H), 4.29 (t, J =5.6 Hz, 1H), 4.12 (m, 2H), 3.92 (s, 3H), 3.86 (s, 3H), 3.08 (m, 1H), 2.97 (m, 4H), 2.80 (s, 1H), 2.57-2.49 (m, 2H), 2.01 (m, 4H), 1.63 (m, 1H), 1.43 (m, 4H), 1.34 (m, 8H), 1.17 (m, 2H).

α-NVOC-ε-(6-((6-((biotinoyl)amino)hexanoyl)amino)hexanoyl)-L-lysine (NVOC-B) Fluorenylmethyl Ester.

A mixture of 6-((6-((biotinoyl)amino)hexanoyl)amino)hexanoic acid, succinimidyl ester (50 mg, 0.088 mmol, 1 eq.) and α-NVOC-lysine fluorenylmethyl ester (75 mg, 0.13 mmol, 1.5 eq.) [52] was stirred in dry DMF (1 mL) with triethylamine (25 μL, 2 eq.). The reaction was stirred for four days after which time the volatiles were removed and the yellow residue was purified by column chromatography (silica, 4:1 CH_2Cl_2:MeOH) to afford 46 mg of the desired product (47% yield): ^{1}H NMR (500 MHz, DMSO-d_6) δ 7.92-7.84 (m, 3H), 7.69-7.55 (m, 6 H), 7.36 (q, J =7.6 Hz, 2H), 7.25 (m, 2H), 7.13 (s, 1H), 6.35 (s, 1H), 6.30 (s, 1H), 5.34 (AB, J =13.5 Hz, 1H), 5.28 (AB, J =13.5 Hz, 1H), 4.45 (m, 1H), 4.38 (m, 1H), 4.23 (m, 1H), 4.19 (t, J=5.0 Hz, 1H), 4.07 (m, 1H), 3.89 (q, J=7.0 Hz, 1H), 3.84 (s, 3H), 3.80 (s, 3H), 3.05 (m, 1H), 2.75 (m, 1H), 1.97 (m, 6H), 1.65- 1.20 (m, 26H); ^{13}C NMR (125 MHz, DMSO-d_6) δ 172.0, 171.6, 171.5, 162.4, 155.5, 153.2, 147.5, 144.0, 143.4, 140.5, 139.0, 127.5, 127.4, 126.8, 124.7, 119.8, 110.3, 108.0, 65.5, 62.2, 60.8, 59.0, 55.9, 55.9, 55.1, 53.7, 46.1, 40.3, 38.0, 37.8, 35.1, 35.0, 29.9, 28.7, 28.4, 27.9, 27.7, 25.9, 25.0, 24.8, 22.5.

α-NVOC-ε-(6-((6-((biotinoyl)amino)hexanoyl)amino)hexanoyl)-L-lysine Cyanomethyl Ester.

α-NVOC-ε-(6-((6-((biotinoyl)amino)hexanoyl)amino)hexanoyl)- L-lysine fluorenylmethyl ester (19 mg, 0.02 mmol, 1 eq.) was dissolved in dry DMF (1 mL). Piperidine (20 μL, 10 eq.) was added and the mixture was stirred. After 2.5 hours, volatiles were removed in vacuo, hexane (15 mL) was added, and the suspension was filtered. The remaining solid was dissolved in aqueous DMF. The solution was lyophilized and the solid residue was dissolved in dry DMF (4 mL), chloroacetonitrile (0.8 mL) and triethylamine (100 μL). After 24 hours, the mixture was separated using preparative HPLC with a linear gradient of 100% 10 mM aqueous acetic acid to 100% acetonitrile over one hour. The product-containing fractions were lyophilized to give 5 mg of a pale yellow solid (31% yield for 2 steps): ^{1}H NMR (500 MHz, DMSO-d_6) δ 8.08 (d, J=7.5 Hz, 1H), 7.95 (s, 1H), 7.18 (s, 1H), 6.40 (s, 1H), 6.34 (s, 1H), 5.40 (AB, J =14.8 Hz, 1H), 5.35 (AB, J =14.8 Hz, 1H), 5.01 (s, 2H), 4.31 (m, 1H), 4.12 (m, 2H), 3.92 (s, 3H), 3.86 (s, 3H), 3.07 (m, 1H), 2.98 (m, 6 H), 2.88 (s, 1H), 2.80 (dd, J=7.3, 5.2Hz, 1H), 2.72 (s, 1H), 2.57 (d, J=12.4 Hz, 1H), 1.75- 1.52 (m, 4H), 1.50-1.41 (m, 6H), 1.41-1.25 (m, 8H), 1.25-1.12 (m, 4H); ^{13}C NMR (125 MHz, DMSO-d_6) δ 171.8, 171.7, 171.3, 162.5, 162.2, 155.6, 153.4, 148.6, 139.3, 127.5, 115.5, 110.5, 108.3, 66.9, 62.5, 61.0, 59.1, 56.1, 56.1, 55.2, 53.5, 49.7, 49.3, 44.2, 38.2, 37.9, 35.6, 35.3, 35.1, 30.7, 28.8, 28.5, 28.0, 27.9, 26.0, 25.2, 24.9, 24.8, 22.6.

6.6.2 General Method for Coupling of N-Protected Amino Acids to dCA.

The N-protected amino acid (≈ 30 μmol, 3 eq.) was mixed with the tetra-*n*-butylammonium salt of the dCA dinucleotide (≈ 10 μmol, 1 eq.) in dry DMF (400 μL). The reactions were monitored by analytical HPLC with a linear gradient of 100% 25 mM aqueous ammonium acetate buffer (pH 4.5) to 100% acetonitrile over one hour. Following the disappearance of the dCA (5 min. to 3 hours), the reaction mixture was separated using semi-preparative HPLC with a linear gradient of 100% 25

mM aqueous ammonium acetate buffer (pH 4.5) to 100% acetonitrile. The fractions containing the aminoacyl dinucleotide were collected and lyophilized. The lyophilized solid was redissolved in 10 mM aqueous acetic acid/acetonitrile and lyophilized a second time to remove salts. The products were characterized by UV-Vis and mass spectrometry.

dCA-NVOC-Biocytin

Prepared as above: FAB-MS $[M\text{-}H]^-$ 1228, calcd for $C_{45}H_{61}N_{13}O_{22}P_2S$ 1228; UV λ_{max} (10 mM Aq. HOAc) 258 nm, 350 nm.

dCA-α-NVOC-ε-(6-(biotinoyl)amino)hexanoyl-L-lysine

Prepared as above: FAB-MS $[M\text{-}H]^-$1342, calcd for $C_{51}H_{72}N_{14}O_{23}P_2S$ 1342; UV λ_{max} (10 mM Aq. HOAc) 256 nm, 350 nm.

dCA-α-NVOC-ε- (6-((6-((biotinoyl)amino)hexanoyl)amino)hexanoyl)-L- lysine

Prepared as above: ESI-MS $[M\text{-}Na]^+$ 1480, calcd for $C_{57}H_{80}N_{15}O_{24}P_2SNa$ 1480; UV λ_{max} (10 mM Aq. HOAc) 253 nm, 350 nm.

6.6.3 Site-Directed Mutagenesis of the AChR α67-76 TAG Mutants

The mouse muscle AChR α67-76 TAG mutants were prepared using a standard two step PCR-based cassette mutagenesis procedure using *Pfu* Polymerase (Stratagene). For each mutant, a trimmed *BsmB*I-*Msc*I PCR-generated fragment was subcloned into the dephosphorylated *BsmB*I-*Msc*I fragment of mouse muscle AChR α subunit in the pAMV-PA vector [30]. When possible, silent mutations were introduced to facilitate screening of positive colonies by restriction endonuclease digestion. We have found the *Bfa*I recognition sequence CTAG to be particularly helpful for screening.

All mutants were verified by automated sequencing of the entire PCR cassette across both ligation sites.

6.6.4 In Vitro Transcription of mRNA

mRNA was synthesized in vitro from *Not*I linearized plasmid DNA (1 μg/40 μL reaction) using the T7 MessageMachine kit from Ambion per the manufacturer's instructions. The products were purified by phenol:chloroform:isoamyl alcohol extraction and isopropanol precipitation. Purity of the mRNA was assayed by gel electrophoresis.

6.6.5 Oocyte Preparation and Injection

Oocytes were removed from *Xenopus laevis* as described previously [40] and maintained at 18 °C in ND96 solution consisting of 96 mM NaCl, 2 mM KCl, 1 mM $MgCl_2$, 1.8 mM $CaCl_2$, 5 mM HEPES (pH 7.5), supplemented with sodium pyruvate (2.5 mM), gentamicin (50 μg/mL), theophylline (0.6 mM) and horse serum (5%). Prior to microinjection, the NVOC-aminoacyl-tRNA (1 μg/μL in 1 mM NaOAc, pH 5.0) was deprotected by irradiating the sample for 5 minutes with a 1000W xenon arc lamp (Oriel) operating at 600W equipped with WG-335 and UG-11 filters (Schott). The deprotected aminoacyl-tRNA was mixed 1:1 with a water solution of the desired mRNA. Oocytes were injected with 50 nL of a mixture containing 25 ng aminoacyl-tRNA and 1.2 ng of total mRNA (4:1:1:1 α:β:γ:δ subunit stoichiometry).

6.6.6 Electrophysiology

Two-electrode voltage clamp measurements were performed 36–48 hours after injection using a GeneClamp 500 amplifier (Axon Instruments, Foster City, CA). Microelectrodes were filled with 3M KCl and had resistances of 0.5 to 2 MΩ. Oocytes were perfused with a nominally calcium-free bath solution containing 96 mM NaCl, 2 mM KCl, 1 mM $MgCl_2$, 5 mM HEPES (pH 7.5). Macroscopic ACh-induced currents were

recorded in response to application of 200 μM ACh at a holding potential of -80 mV. Whole-cell current measurements are reported as the mean $\pm$ s.e.m.

6.6.7 General Procedure for Radioactive Ligand Binding Experiments

Batches of five oocytes were incubated at room temperature for 15 min with gentle agitation in 400 μL calcium-free ND96 (see Electrophysiology section) containing 10 mg/mL BSA (FractionV, Sigma). The radioactive ligand was then added to a final concentration of 300 pM for [^{125}I]-streptavidin (Amersham, 30 μCi/μg) or 1 nM for [^{125}I]-α-bungarotoxin (Amersham, 2000 Ci/mmol). After two hours, the oocytes were transferred to an ice-cold solution of ND96 + 10 mg/mL BSA (1.5 mL) and agitated for five min. The oocytes were then washed extensively in fresh solutions of ice-cold ND96 + BSA and transferred individually to scintillation vials and counted in a Beckman LS5000 γ-counter. In all cases, a sample of the final wash solution was counted to ensure that unbound ligand had not been carried through the assay. In addition, five empty scintillation vials were counted in each experiment to determine the detector background which was then subtracted from subsequent measurements. In experiments involving the binding of both labeled and unlabeled ligand, the unlabeled ligand was added at high concentration (166 nM for streptavidin, 312 nM for bungarotoxin) to the initial 400 μL solution. After 20 min, the labeled ligand was added and the assay proceeded as described above.

Bibliography

[1] S. H. White, editor. *Membrane Protein Structure.* Methods in Physiology. Oxford University Press, Inc, New York, 1994.

[2] M. Hollmann, A. O'Shea-Greenfield, S. W. Rogers, and S. Heinemann. Cloning by functional expression of a member of the glutamate receptor family. *Nature*, 342:643–648, 1989.

[3] M. Hollmann, C. Maron, and S. Heinemann. N-glycosylation site tagging suggests a three transmembrane domain topology for the glutamate receptor GluR1. *Neuron*, 13:1331–1343, 1994.

[4] J. A. Bennett and R. Dingledine. Topology profile for a glutamate receptor: Three transmembrane domains and a channel-lining reentrant membrane loop. *Neuron*, 14:373–384, 1995.

[5] R. A. Schwalbe, Z. Wang, L. Bianchi, and A. M. Brown. Novel sites of N-glycosylation in ROMK1 reveal the putative pore-forming segment H5 as extracellular. *J. Biol. Chem.*, 271:24201–24206, 1996.

[6] E. R. Bennett and B. I. Kanner. The membrane topology of GAT-1, a (Na^+ + Cl^-)-coupled γ-aminobutyric acid transporter from rat brain. *J. Biol. Chem.*, 272:1203–1210, 1997.

[7] L. Olivares, C. Aragón, C. Giménez, and F. Zafra. Analysis of the transmembrane topology of the glycine transporter GLYT1. *J. Biol. Chem.*, 272:1211–1217, 1997.

[8] E. Turk, C. J. Kerner, M. P. Lostao, and E. M. Wright. Membrane topology of the human Na^+/glucose cotransporter SGLT1. *J. Biol. Chem.*, 271:1925–1934, 1996.

[9] C. H. Pedemonte, G. Sachs, and J. H. Kaplan. An intrinsic membrane glycoprotein with cytosolically oriented N-linked sugars. *Proc. Natl. Acad. Sci. USA*, 87:9789–9793, 1990.

[10] X. B. Chang, Y. X. Hou, T. J. Jensen, and J. R. Riordan. Mapping of cystic fibrosis transmembrane conductance regulator membrane topology by glycosylation site insertion. *J. Biol. Chem.*, 269:18572–18575, 1994.

[11] S. Frillingos and H. R. Kaback. Probing the conformation of the lactose permease of *Escherichia coli* by *in situ* site-directed sulfhydryl modification. *Biochemistry*, 35:3950–3956, 1996.

[12] T. W. Loo and D. M. Clarke. Membrane topology of a cysteine-less mutant of human P-glycoprotein. *J. Biol. Chem.*, 270:843–848, 1995.

[13] J. Sun, J. Li, N. Carrasco, and H. R. Kaback. The last two cytoplasmic loops in the lactose permease of *Escherichia coli* comprise a discontinuous epitope for a monoclonal antibody. *Biochemistry*, 36:274–280, 1997.

[14] S. L. Slatin, X. Q. Qiu, K. S. Jakes, and A. Finkelstein. Identification of a translocated protein segment in a voltage-dependent channel. *Nature*, 371:158–161, 1994.

[15] M.H. Akabas, D.A. Stauffer, M. Xu, and A. Karlin. Acetylcholine receptor channel structure probed in cysteine-substitution mutants. *Science*, 258:307–310, 1992.

[16] J. A. Javitch, D. Fu, J. Chen, and A. Karlin. Mapping the binding-site crevice of the dopamine D2 receptor by the substituted-cysteine accessibility method. *Neuron*, 14:825–831, 1995.

[17] R. A. Chavez and Z. W. Hall. Expression of fusion proteins of the nicotinic acetylcholine receptor from mammalian muscle identifies the membrane-spanning regions in the α and δ subunits. *J. Cell Biol.*, 116:385–393, 1992.

[18] R. Anand, L. Bason, M. S. Saedi, V. Gerzanich, X. Peng, and J. Lindstrom. Reporter epitopes: A novel approach to examine transmembrane topology of integral membrane proteins applied to the α subunit of the nicotinic acetylcholine receptor. *Biochemistry*, 32:9975–9984, 1993.

[19] M. Mund, C. Weise, P. Franke, and F. Hucho. Mapping of exposed surfaces of the nicotinic acetylcholine receptor by identification of iodinated tyrosine residues. *J. Prot. Chem.*, 16:161–170, 1997.

[20] N. M. Green. Avidin. *Advan. Protein Chem.*, 29:85–133, 1975.

[21] M. Wilchek and E. A. Bayer, editors. *Avidin-Biotin Technology*, volume 184 of *Methods in Enzymology*. Academic Press, Inc., San Diego, CA, 1990.

[22] K. Hofmann, G. Titus, J. A. Montibeller, and F. M. Finn. Avidin binding of carboxyl-substituted biotin and analogues. *Biochemistry*, 21:978–984, 1982.

[23] T. Kurzchalia, M. Wiedmann, H. Breter, W. Zimmermann, E. Bauschke, and T. A. Rapoport. tRNA-Mediated labelling of proteins with biotin. A nonradioactive method for the detection of cell-free translation products. *Eur. J. Biochem.*, 172:663–668, 1988.

[24] C. J. Noren, S. J. Anthony-Cahill, M.C. Griffith, and P.G. Schultz. A general method for site-specific incorporation of unnatural amino acids into proteins. *Science*, 244:182–188, 1989.

[25] J. Ellman, D. Mendel, S. Anthony-Cahill, C. J. Noren, and P. G. Schultz. Biosynthetic method for introducing unnatural amino-acids site-specifically into proteins. *Meth. Enzym.*, 202:301–336, 1991.

[26] S. A. Robertson, J. A. Ellman, and P. G. Schultz. A general and efficient route for chemical aminoacylation of transfer RNAs. *J. Am. Chem. Soc.*, 113:2722–2729, 1991.

[27] V. W. Cornish, D. Mendel, and P. G. Schultz. Probing protein structure and function with an expanded genetic code. *Angew. Chem. Int. Ed. Engl.*, 34:621–633, 1995.

[28] M. W. Nowak, P. C. Kearney, J. R. Sampson, M. E. Saks, C. G. Labarca, S. K. Silverman, W. Zhong, J. Thorson, J. N. Abelson, N. Davidson, P. G. Schultz, D. A. Dougherty, and H. A. Lester. Nicotinic receptor binding site probed with unnatural amino-acid incorporation in intact cells. *Science*, 268:439–442, 1995.

[29] M. E. Saks, J. R. Sampson, M. W. Nowak, P. C. Kearney, F. Du, J. N. Abelson, H. A. Lester, and D. A. Dougherty. An engineered tetrahymena $tRNA^{Gln}$ for *in vivo* incorporation of unnatural amino acids into proteins by nonsense suppression. *J. Biol. Chem.*, 271:23169–23175, 1996.

[30] M. W. Nowak, J. P. Gallivan, S. K. Silverman, C. G. Labarca, D. A. Dougherty, and H. A. Lester. *In vivo* incorporation of unnatural amino acids into ion channels in a *Xenopus* oocyte expression system. *Meth. Enzymol.*, 293:504–529, 1998.

[31] A. Karlin. Structure of nicotinic acetylcholine receptors. *Curr. Opin. Neurobiol.*, 3:299–309, 1993.

[32] A. Devillers-Thiéry, J. L. Galzi, J. L. Eiselé, S. Bertrand, D. Bertrand, and J. P. Changeux. Functional architecture of the nicotinic acetylcholine receptor: A prototype of ligand-gated ion channels. *J. Membrane Biol.*, 136:97–112, 1993.

[33] H. A. Lester. The permeation pathway of neurotransmitter-gated ion channels. *Annu. Rev. Biophys. Biomol. Struct.*, 21:267–292, 1992.

[34] R.M. Stroud, M.P. McCarthy, and M. Shuster. Nicotinic acetylcholine receptor superfamily of ligand-gated ion channels. *Biochemistry*, 29:11009–11023, 1990.

[35] N. Unwin. Nicotinic acetylcholine receptor at 9 Å resolution. *J. Mol. Biol.*, 229:1101–1124, 1993.

[36] E. Kubalek, S. Ralston, J. Lindstrom, and N. Unwin. Location of subunits within the acetylcholine receptor by electron image analysis of tubular crystals from *Torpedo marmorata*. *J. Cell Biol.*, 105:9–18, 1987.

[37] S. J. Tzartos, M. T. Cung, P. Demange, H. Loutrari, A. Mamalaki, M. Marraud, I. Papadouli, C. Sakarellos, and V. Tsikaris. The main immunogenic region (MIR) of the nicotinic acetylcholine receptor and the anti-MIR antibodies. *Molecul. Neurobiol.*, 5:1–29, 1991.

[38] A. Mamalaki and S. J. Tzartos. Nicotinic acetylcholine receptor: Structure, function and main immunogenic region. *Adv. Neuroimmunol.*, 4:339–354, 1994.

[39] R. Beroukhim and N. Unwin. Three-dimensional location of the main immunogenic region of the acetylcholine receptor. *Neuron*, 15:323–331, 1995.

[40] M. Quick and H. A. Lester. *Methods for Expression of Excitablitity Proteins in Xenopus Oocytes*. Academic Press, San Diego, CA, 1994.

[41] S. Verrall and Z. W. Hall. The N-terminal domains of acetylcholine receptor subunits contain recognition signals for the initial steps of receptor assembly. *Cell*, 68:23–31, 1992.

[42] S. Tamamizu, D. H. Butler, J. A. Lasalde, and M. G. McNamee. Effects of antibody binding on structural transitions of the nicotinic acetylcholine receptor. *Biochemistry*, 35:11773–11781, 1996.

[43] V. Tsikaris, E. Detsikas, M. Sakarellos-Daitsiotis, C. Sakarellos, E. Vatzaki, S. Tzartos, M. Marraud, and M. T. Cung. Conformational requirements for molecular recognition of acetylcholine receptor main immunogenic region (MIR) analogues by monoclonal anti-MIR antibody: A two-dimensional nuclear magnetic resonance and molecular dynamics approach. *Biopolymers*, 33:1123–1134, 1993.

[44] E. E. Abola, F. C. Bernstein, S. H. Bryant, T. F. Koetzle, and J. Weng. *Protein Data Bank.* Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester, 1987.

[45] P. C. Weber, D. H. Ohlendorf, J. J. Wendoloski, and F. R. Salemme. Structural origins of high-affinity biotin binding to streptavidin. *Science*, 243:85–88, 1989.

[46] K. Fujita and J. Silver. Surprising lability of biotin-streptavidin bond during transcription of biotinylated DNA bound to paramagnetic streptavidin beads. *BioTechniques*, 14:608–617, 1993.

[47] V.W. Cornish, K.M. Hahn, and P.G. Schultz. Site-specific protein modification using a ketone handle. *J. Am. Chem. Soc.*, 118(34):8150–8151, 1996.

[48] L.K. Mahal, K.J. Yarema, and C.R. Bertozzi. Engineering chemical reactivity on cell surfaces through oligosaccharide biosynthesis. *Science*, 276(5315):1125–1128, 1997.

[49] E. Saxon and C.R. Bertozzi. Cell surface engineering by a modified Staudinger reaction. *Science*, 287(5460):2007–2010, 2000.

[50] P. C. Kearney, M. W. Nowak, W. Zhong, S. K. Silverman, H. A. Lester, and D. A. Dougherty. Dose-response relations for unnatural amino acids at the agonist binding site of the nicotinic acetylcholine receptor: Tests with novel side chains and with several agonists. *Mol. Pharmacol.*, 50:1401–1412, 1996.

[51] B. Amit, U. Zehavi, and A. Patchornik. Photosensitive protecting groups of amino sugars and their use in glycoside synthesis. 2-Nitrobenzyloxycarbonylamino and 6-nitroveratryloxycarbonylamino derivatives. *J. Org. Chem.*, 39:192–196, 1974.

[52] S. K. Silverman. *I. Conformational and Charge Effects on High-Spin Organic Polyradicals; II. Studies on the Chemical-Scale Origin of Ion Selectivity in Potassium Channels.* Ph.D. Thesis, California Institute of Technology, 1998.

Chapter 7 Other Experiments

7.1 Amino Acid Synthesis

In this section, the synthesis of a number of natural and unnatural amino acids is outlined. The syntheses generally occur in three stages:

1. The α-NH_2 group and other reactive side chain groups are protected
2. The protected amino acid is activated as the cyanomethyl ester
3. The activated ester is coupled to the dCA-dinucleotide

General methods describing these steps are outlined below.

For simple amino acids with unreactive side chains, only the α-NH_2 group needs to be protected. Protection of the α-NH_2 group increases the stability of the aminoacyl-dinucleotide and aminoacyl-tRNAs. Prior to use, the aminoacyl tRNAs must be deprotected. The conditions used to deprotect the amino acid must be mild, such that neither the amino acid, the tRNA, nor the oocytes are damaged. Originally, the *o*-nitroveratryloxycarbonyl (NVOC) protecting group was used becuase it could be removed under mild conditions using 350 nm light. However, in some applications, it is desirable to have an alternate protecting group for the α-amine. For example, it is possible to produce "caged" amino acids that have side chains that are quickly unmasked at a specified time. Photochemical decaging offers the ability to spatially and temporally reveal new functionality within a protein. However, because the 350 nm used to remove the NVOC α-amine protecing group would likely also remove the side chain caging group, new α-amine protecting groups had to be adapted for use that were compatible with both aminoacyl-tRNAs and the *Xenopus* oocyte.

Two protecting groups were investigated as potential complements to the NVOC group. The 4-pentenoyl group (4-PO), originally developed by Fraser-Reid and co-workers is normally removed by a solution of I_2 in aqueous THF [1]. To test whether this protecting group was compatible with the oocyte system, the N-4-PO-phenylalanine was synthesized, coupled to the dCA dinucleotide, and ligated to tRNA-THG73(-CA). The protected aminoacyl-tRNA was treated with 1 volume of a saturated solution of I_2 in 1 mM NH_4OAc, pH 4.5, for five minutes, combined with

mRNA encoding the nAChR (1.1 ng total/oocyte, 4:1:1:1 ratio of αE180TAG, β, γ, δ subunits), and injected into *Xenopus* oocytes. After 24 hours, whole-cell currents were robust, indicating that the protecting group had been sufficiently removed.

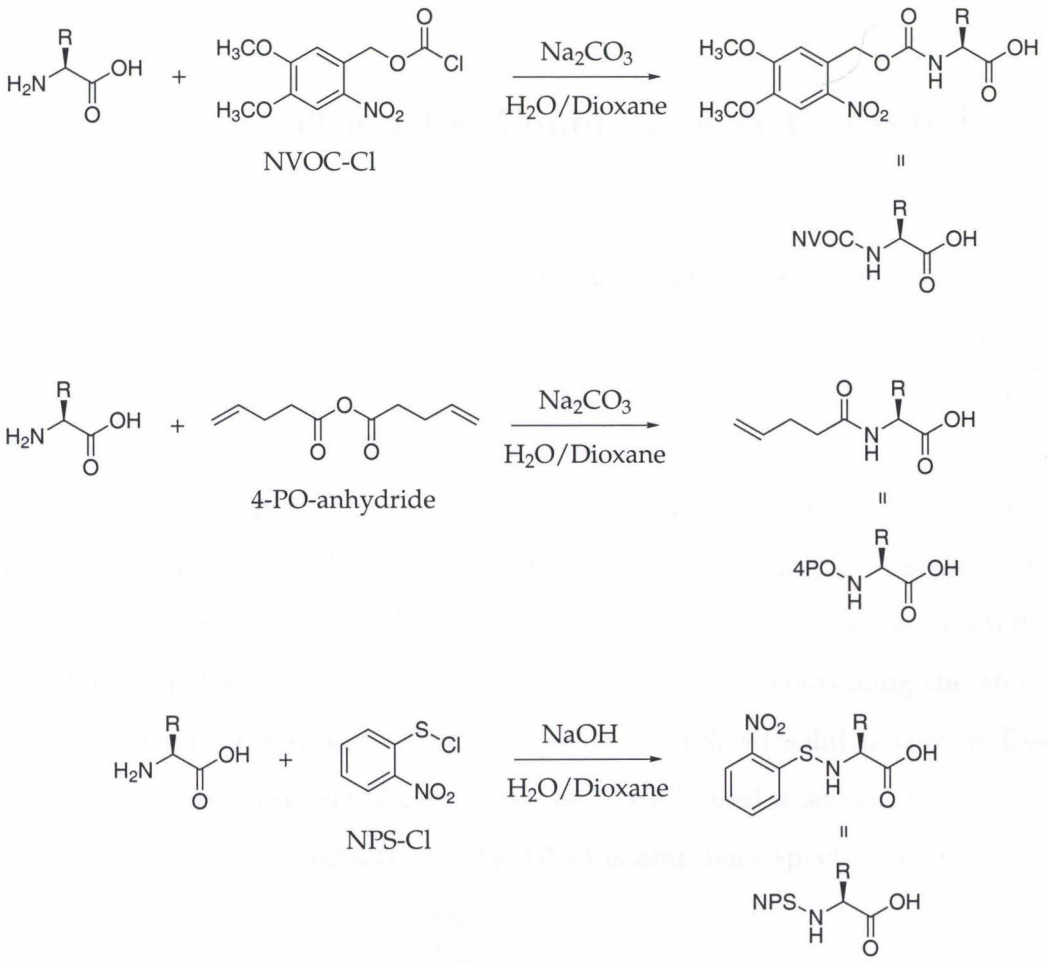
In addition to the 4-PO group, the *o*-nitrophenyl sulfenyl (NPS) group was also investigated [2]. The NPS group is normally removed by reducing reagents, such as sodium thiosulfate. Thus, NPS-phenylalanine was synthesized, coupled to the dCA dinucleotide, and ligated to tRNA-THG73(-CA). The protected aminoacyl-tRNA (12.5 ng/oocyte) was treated with 20 mM $Na_2S_2O_3$ in 1 mM NH_4OAc, for 10 minutes, combined with mRNA encoding the nAChR (1.1 ng total/oocyte, 4:1:1:1 ratio of αE180TAG, β, γ, δ subunits), and injected into *Xenopus* oocytes. After 24 hours, whole-cell currents in response to 200 μM ACh were robust ($>$ 1 μA), indicating that the protecting group had been removed. Since the interior of the oocyte is a reducing environment, it was believed that deprotection prior to injection might be unnecessary. Preliminary results indicate that this is the case. In cases where aminoacyl-tRNA stability is critical, this group may provide a means to keep the tRNA protected (and thus stable) for longer periods of time. However, NPS amino acids are somewhat less attractive synthetically than those protected with either the NVOC or the 4-PO group.

7.1.1 General Method for the Protection of Amino Acids

Scheme 7.1 outlines the syntheses of N-protected amino acids using the NVOC, 4-PO, and NPS protecting groups. For the NVOC and 4-PO groups, Schotten-Baumann conditions are used to install the protecting group.[1] The amino acid (1.0 eq.) and Na_2CO_3 (2.0 eq.) are dissolved in a 1:1 mixture of water/dioxane. A solution of either NVOC-Cl (1.0 eq.) [3] or 4-PO-anhydride (1.1–1.3 eq.) [4] in dioxane is added and the mixture is stirred until TLC indicates that the reaction is complete (usually 1-4 hours). Reactions involving 4-PO-anhydride may be sluggish and additional 4-PO-anhydride may be added. When the reaction is judged to be complete by TLC,

[1]For a procedure to install the NPS group, see Section 7.2.1.

CH_2Cl_2 is added, followed by an equal amount of 1 N $NaHSO_4$ and the aqueous layer is extracted three times with CH_2Cl_2. The organic extracts are combined, dried over Na_2SO_4, and concentrated in vacuo. The crude products are purified by flash chromatography on silica. In many cases, a mixture of hexane/ethyl acetate will elute the byproducts, and switching to ethyl acetate with 1% acetic acid will elute the protected amino acid.



Scheme 7.1: Protection of amino acids.

7.1.2 General Method for the Synthesis of Amino Acid Cyanomethyl Esters

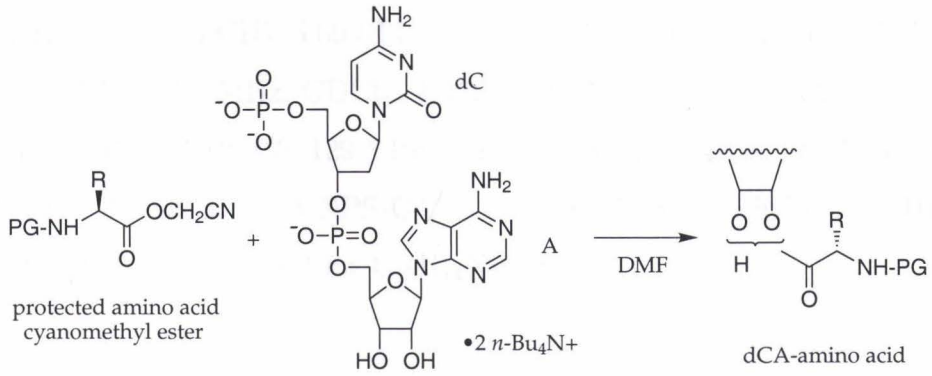
The protected amino acid ($\approx$ 0.5 mmol, 1.0 eq.) is dissolved in 2 mL dry DMF under argon (Scheme 7.2). Triethylamine (3 eq.) is added, followed by 2 mL chloroace-

tonitrile. The reactions are stirred under argon for 1–24 hours and the volatiles are removed in vacuo. The products are purified by flash chromatography. Useful solvent systems are CH_2Cl_2, or 1:1 hexane/ethyl acetate.

Scheme 7.2: Cyanomethylation of protected amino acids.

7.1.3 General Method for Coupling of N-Protected Amino Acids to dCA

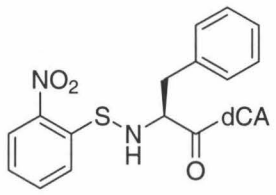
This procedure is adapted from previously published procedures [5–7]. The N-protected amino acid ($\approx$ 30 μmol, 3 eq.) is mixed with the tetra-*n*-butylammonium salt of the dCA dinucleotide ($\approx$ 10 μmol, 1 eq.) in dry DMF (400 μL). The reactions are monitored by analytical HPLC with a linear gradient of 100% 25 mM aqueous ammonium acetate buffer (pH 4.5) to 100% acetonitrile over one hour. Following the disappearance of the dCA (5 min. to 3 hours), the reaction mixture is separated using semi-preparative HPLC with a linear gradient of 100% 25 mM aqueous ammonium acetate buffer (pH 4.5) to 100% acetonitrile. The fractions containing the aminoacyl dinucleotide are collected and lyophilized. The lyophilized solid is then redissolved in 10 mM aqueous acetic acid/acetonitrile and lyophilized a second time to remove salts. The products are characterized by UV-Vis and mass spectrometry.



Scheme 7.3: Synthesis of dCA amino acids.

7.2 Synthesis of Natural Amino Acids

7.2.1 dCA-NPS-Phenylalanine



NPS-Phenylalanine

Phenylalanine (330 mg, 2.0 mmol) was dissolved in dioxane (20 mL), water (5 mL) and 1.14 M NaOH (1.75 mL, 1 eq). Subsequently, 2-nitrophenylsulfenyl chloride (417 mg, 2.2 mmol, 1.1 eq) was added in small portions, turning the yellow mixture dark green. Addition of 1.93 mL of 1.14 M NaOH returned the mixture to its original yellow color. After 30 minutes, CH_2Cl_2 (25 mL) was added, followed by $NaHSO_4$ (1 N, 25 mL). The mixture was extracted 3 times with CH_2Cl_2 (25 mL). The combined organic layers were dried over $MgSO_4$, filtered and the solvent was evaporated to give a yellow solid. The crude product was purified by flash chromatography (silica, 50:50 hexane/ethyl acetate followed by 1% HOAc in ethyl acetate) to give the protected amino acid as a yellow solid (503 mg, 79% yield): ^{1}H NMR (300 MHz, $CDCl_3$) δ 11.35 (br s, 1 H, COOH), 8.21 (dd, J = 7.0, 1.2 Hz, 1 H, NPS-3-H), 7.44–7.20 (m, 7 H), 7.17 (td, J = 5.5, 1.5 Hz, 1 H, Phe-4-H), 3.80 (m, 1 H, α-CH), 3.34 (dd, J =14.0, 4.0 Hz, 1 H, side chain CH), 3.00 (dd, J = 14.0, 4.0 Hz, 1 H, side chain CH), 3.20 (br s, 1 H); ^{13}C NMR (75 MHz, $CDCl_3$) δ 178.9 COOH, 144.3 NPS-C2, 142.5 NPS-C1, 136.4 Phe-C1, 133.8 NPS-C5, 129.5 Phe-C2,C6, 128.8 Phe-C3,5, 127.3 NPS-C6, 125.6 Phe-C4, 124.7 NPS-C3, 124.4 NPS-C4, 65.5 α-CH, 39.6 side chain CH_2; HRMS CI$^+$ (NH_3) [M-H]$^+$ 319.0757, calcd for $C_{15}H_{14}N_2O_4S$ 319.0753.

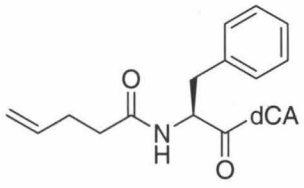
NPS-Phenyalanine Cyanomethyl Ester

NPS-phenylalanine (100 mg, 0.31 mmol) was dissolved in 2 mL dry DMF under argon. Triethylamine (0.13 mL, 3 eq.) was added, followed by chloroacetonitrile (2 mL). After 5 hours, the volatiles were removed in vacuo to give a yellow oil. The product was purified by flash chromatography (silica, hexane, 80:20 hexane/ethyl acetate, 50:50 hexane/ethyl acetate) to give a yellow oil (104 mg, 93%): ^{1}H NMR (300 MHz, $CDCl_3$) δ 8.18 (dd, J = 7.0, 1.0 Hz, 1 H, NPS-3-H), 7.45–7.20 (m, 7 H), 7.18 (td, J = 5.6, 1.2 Hz, 1 H, Phe-H-4), 4.82 (dd, J = 15.6, 12.0 Hz, 2 H, OCH2), 3.82 (m, 1 H, α-CH), 3.35-3.25 (overlapping peaks, 2 H, side chain CH, and NH), 3.03 (m, 1 h, side chain CH); ^{13}C NMR (75 MHz, $CDCl_3$) δ 171.8 COOR, 143.8 NPS-C2, 142.4 NPS-C1, 135.7 Phe-C1, 133.8 NPS-C5, 129.4 Phe-C2,C6, 128.8 Phe-C3,5, 127.3 NPS-C6, 125.6 Phe-C4, 124.8 NPS-C3, 124.4 NPS-C4, 113.80 CN, 65.4 α-CH, 48.9 OCH_2, 39.5 side chain CH_2; HRMS EI$^+$ [M-H]$^+$ 357.0779, calcd for $C_{17}H_{15}N_3O_4S$ 357.0783.

dCA-NPS-Phenylalanine

Prepared as above: FAB-MS [M-H]$^-$ 935, calcd for $C_{34}H_{38}N_{10}O_{16}P_2S$ 935; UV λ_{max} (10 mM Aq. HOAc) 251 nm, 384 nm.

7.2.2 dCA-4-PO-Phenylalanine



4-PO-Phenylalanine

A solution of phenylalanine (100 mg, 0.6 mmol) and sodium carbonate (158 mg, 0.6 mmol) in 15 mL water and 7 mL dioxane was cooled in an ice bath. A solution of

4-pentenoic anhydride (130 mg, 0.72 mmol, 1.2 eq) in dioxane was added dropwise. After 24 hours, 50 mL 1 N $NaHSO_4$ was added, followed by 40 mL CH_2Cl_2. The mixture was extracted three times with 40 mL portions of CH_2Cl_2. The organic layers were combined, dried over $MgSO_4$ and concentrated in vacuo to give 175 mg of crude product that crystallized in an NMR tube. The remainder was used without further purification: ^{1}H NMR (300 MHz, $CDCl_3$) δ 10.36 (s, 1 H, COOH), 7.25 (m, 3 H, Ar-H), 7.16 (dd, J = 6.2, 1.7 Hz, 2 H, Ar-H), 6.50 (d J = 7.7 Hz, 1 H, NH), 5.90-5.65 (m, 1 H, CH_2=CH), 5.12- 4.86 (m, 3 H, CH_2=CH and α-CH), 3.16 (dd, J = 6.0, 5.5 Hz., 2 H, side chain CH_2), 2.47-2.26 (m, 4 H, CH_2's in 4-PO group); ^{13}C NMR (75 MHz, $CDCl_3$) δ 178.4 COOH, 174.7 CONH, 136.5 CH_2=CH, 135.7 Ar-C1, 129.4 Ar-C2,6, 128.5 Ar-C3,5, 127.1 Ar-C4, 115.8 CH_2=CH, 53.2 α-CH, 37.2 side chain CH_2, 35.4, 29.4 CH_2's in 4-PO group.

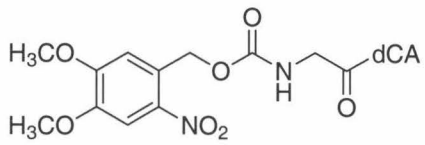
4-PO-Phenylalanine Cyanomethyl Ester

To a solution of 4-pentenoyl phenylalanine (59 mg, 0.24 mmol) in 2 mL dry DMF was added triethylamine (100 μL, 0.72 mmol, 3 eq.) and 2 mL chloroacetonitrile. After 10 hours, the volatiles were removed in vacuo to afford a crude white residue that was purified by flash chromatography (silica, hexane→3:1 hexane/ethyl acetate→1:1 hexane/ethyl acetate) to afford 51 mg of a white solid (75% yield): ^{1}H NMR (300 MHz, $CDCl_3$) δ 7.36-7.25 (m, 3 H, Ar-H), 7.14 (dd, J = 6.2, 1.6 Hz, 2 H, Ar-H), 6.07 (d J = 7.5 Hz, 1 H, NH), 5.85–5.68 (m, 1 H, CH_2=CH), 5.12–5.04 (m, 3 H, CH_2=CH and α-CH), 4.91 (dd, J = 32.3, 15.7 Hz., 2 H, OCH_2CN), 3.20–3.06 (m, 2 H, side chain CH_2) 2.47-2.26 (m, 4 H, CH_2's in 4-PO group); ^{13}C NMR (75 MHz, $CDCl_3$) δ 178.4 COOH, 174.7 CONH, 136.5 CH_2=CH, 135.7 Ar-C1, 129.4 Ar-C2,6, 128.5 Ar-C3,5, 127.1 Ar-C4, 115.8 CH_2=CH, 53.2 α-CH, 37.2 side chain CH_2, 35.4, 29.4 CH_2's in 4-PO group.

dCA-4-PO-Phenylalanine

Prepared as above: FAB-MS $[M-H]^-$ 864.3, calcd for $C_{33}H_{41}N_9O_{15}P_2$ 865.2.

7.2.3 dCA-NVOC-Glycine



NVOC-Glycine

Prepared in 85% yield as above starting with glycine (75 mg, 1 mmol), NVOC-Cl (276 mg, 1 mmol) and (212 mg, 2 mmol): ^{1}H NMR (300 MHz, DMSO-d_6) δ 13.25–12.25 (br s, 1 H, COOH), 7.85 (t, J = 6.2 Hz, 1 H, NH), 7.68 (s, 1 H, Ar-H), 7.17 (s, 1 H, Ar-H), 5.35 (s, 2 H, NVOC-CH_2), 3.92 (s, 3 H, CH_3O), 3.85 (s, 3 H, CH_3O), 3.69 (d, J = 6.2 Hz, 2 H, α-CH_2); ^{13}C NMR (75 MHz, DMSO-d_6) δ 172.1 COOH, 156.7 NHCO, 153.96 Ar-C, 148.08 Ar-C, 139.44 Ar-C, 128.65 Ar-C, 110.46 Ar-CH, 108.52 Ar-CH, 63.05 NVOC-CH_2, 56.65 CH_3O, 56.54 CH_3O, 42.60 α-CH_2; HRMS CI$^+$ (NH_3) [M-NH_4]$^+$ 332.1107, calcd for $C_{12}H_{14}N_2O_8$ + NH_3 332.1094.

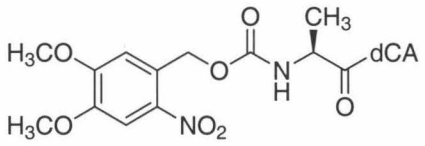
NVOC-Glycine Cyanomethyl Ester

NVOC-Glycine (100 mg, 0.32 mmol) was dissolved in 2 mL dry DMF. Triethylamine (135 μL, 0.96 mmol, 3 eq.) was added, followed by 2 mL chloroacetonitrile. After 6 hours, the volatiles were removed in vacuo and the crude product was purified by flash chromatography (silica, 1:1 hexane/ethyl acetate) to give 112 mg of a pale yellow solid (quantitative yield): ^{1}H NMR (300 MHz, DMSO-d_6) δ 8.07 (t, 1 H, J = 6.2 Hz, NH), 7.67 (s, 1 H, Ar-H), 7.16 (s, 1 H, Ar-H), 5.36 (s, 2 H, NVOC-CH_2), 3.99 (d, J =6.0 Hz, 2 H, α-CH_2), 3.91 (s, 3 H, CH_3O), 3.85 (s, 3 H, CH_3O); ^{13}C NMR (75 MHz, DMSO-d_6) δ 170.0 COOR, 156.6 NHCO, 153.9 Ar-C, 148.2 Ar-C, 139.6 Ar-C, 128.2 Ar-C, 116.3 CN, 110.7 Ar-CH, 108.5 Ar-CH, 63.3 NVOC-CH_2, 56.7 CH_3O, 56.5 CH_3O, 49.9 OCH_2CN, 42.3 α- CH_2; HRMS EI$^+$ [M-H]$^+$ 353.0852, calcd for $C_{14}H_{15}N_3O_8$ 353.0859.

dCA-NVOC-Glycine

Prepared as above: FAB-MS $[M\text{-}H]^-$ 932.2, calcd for $C_{31}H_{38}N_{10}O_{20}P_2$ 932.17.

7.2.4 dCA-NVOC-Alanine



NVOC-Alanine

The reaction was performed as above, and purified by flash chromatography (silica, CH_2Cl_2→5:1 CH_2Cl_2/ethyl acetate→1:1 CH_2Cl_2/ethyl acetate→ethyl acetate →1% acetic acid in ethyl acetate). The product was obtained as an orange/yellow solid in 47% yield: ^{1}H NMR (300 MHz, CD_3OD) δ 7.73 (s, 1 H, Ar-H), 7.19 (s, 1 H, Ar-H), 5.52 (*AB*, J = 15.1 Hz, 1 H, NVOC-CH), 5.40 (*AB*, J = 15.1 Hz, 1 H, NVOC-CH), 4.20 (m, 1 H, α-CH), 3.97 (s, 3 H, CH_3O), 3.90 (s, 3 H, CH_3O), 1.42 (d, J = 7.5 Hz, 3 H, CH_3); ^{13}C NMR (75 MHz, CD_3OD) δ 176.8 COOH, 158.4 CONH, 155.8 Ar-C, 149.9 Ar-C, 141.0 Ar-C, 130.3 Ar-C, 111.1 Ar-CH, 109.7 Ar-CH, 64.9 NVOC-CH_2, 57.4, 57.2 CH_3O's, 51.3 α-CH, 18.2 CH_3; HRMS EI$^+$ $[M\text{-}H]^+$ 328.0914, calcd for $C_{13}H_{16}N_2O_8$ 328.0907.

NVOC-Alanine Cyanomethyl Ester

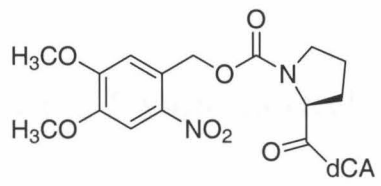
NVOC-alanine (202 mg, 0.62 mmol) was dissolved in 2 mL dry DMF. Triethylamine (250 μL, 1.8 mmol, 3 eq.) was added, followed by 2 mL chloroacetonitrile. After 1.5 hours, the volatiles were removed in vacuo and the crude product was purified by flash chromatography (silica, CH_2Cl_2→95:5 CH_2Cl_2/ethyl acetate) to give 189 mg of pale yellow crystals (84% yield): ^{1}H NMR (300 MHz, $CDCl_3$) δ 7.68 (s, 1 H, Ar-H), 6.98 (s, 1 H, Ar-H), 5.55 (m, 1 H, NH), 5.53 (*AB*, J = 15.1 Hz, 1 H, NVOC-CH),

5.45 (*AB*, J = 15.1 Hz, 1 H, NVOC-CH), 4.79 (m, 2 H, OCH_2CN), 4.20 (m, 1 H, α-CH), 3.99 (s, 3 H, CH_3O), 3.93 (s, 3 H, CH_3O), 1.48 (d, J = 7.5 Hz, 3 H, CH_3); ^{13}C NMR (75 MHz, $CDCl_3$) δ 171.8 COOR, 155.4 CONH, 153.8 Ar-C, 148.2 Ar-C, 139.6 Ar-C, 128.9 Ar-C, 114.1 CN, 109.9 Ar-CH, 108.2 Ar-CH, 64.1 NVOC-CH_2, 56.6, 56.5 CH_3O's, 49.6 α-CH, 49.2 OCH_2CN, 17.9 CH_3; HRMS EI^+ $[M\text{-}H]^+$ 367.1026, calcd for $C_{15}H_{17}N_3O_8$ 367.1016.

dCA-NVOC-Alanine

Prepared as above: FAB-MS $[M\text{-}H]^-$ 945, calcd for $C_{32}H_{40}N_{10}O_{20}P_2$ 945; UV λ_{max} (10 mM Aq. HOAc) 249 nm, 265 nm (sh), 350 nm.

7.2.5 dCA-NVOC-Proline



NVOC-Proline

The reaction was performed as above. The crude product was purified by flash chromatography (silica, 1:1 hexane/ethyl acetate→1% acetic acid in ethyl acetate). Product obtained as a foamy yellow solid in 85% yield. The ^1H NMR spectrum shows 2 poorly resolved conformational isomers: ^1H NMR (300 MHz, CD_3OD) δ 7.59 (s - br, 1 H), (7.00,6.99*) (s, 1 H), H_2O obscures 5.8–5.5 ppm, 5.27 (d, J = 23 Hz, 1 H), 5.22 (d, J = 23 Hz, 1 H), (4.48,4.44)(dd, J = 8.0, 3.5 Hz, 1 H), 4.36 (dd, J = 8.0, 3.5 Hz, 1 H), (3.92,3.91*) (s, 3 H), (3.84,3.83*) (s, 3 H), 3.65–3.40 (m, 2 H), 2.42–1.90 (m - br); ^{13}C NMR (75 MHz, CD_3OD) δ (175.9, 175.2*), (155.9, 155.6*), (155.3, 155.0*), (149.2, 149.0*), (140.2, 139.9*), (110.3, 110.0*), (108.9, 108.8*), (65.3, 65.1*), (60.5, 60.1*), 57.25, (56.8 (2)), (56.7 (2)), (48.2, 47.8*), (31.9, 30.9*), (25.2, 24.3*).

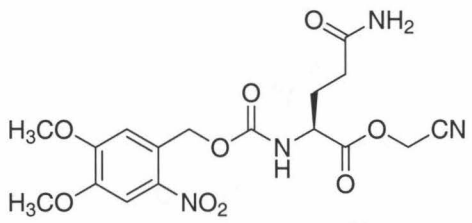
NVOC-Proline Cyanomethyl Ester

The reaction was performed as above. The crude product was purified by flash chromatography (silica, 1:1 hexane/ethyl acetate). Product obtained as a yellow solid in 80% yield. The ^{1}H NMR spectrum shows 2 poorly resolved conformational isomers: ^{1}H NMR (300 MHz, $CDCl_3$) δ (7.71, 7.69*) (s, 1 H), (7.01, 6.98*) (s, 1 H), 5.77–5.38 (m, 2 H), 4.93–4.68 (m, 2 H), 4.56–4.42 (m, 1 H), (4.04, 4.02*) (s, 3 H), (3.96, 3.95*) (s, 3 H), 3.70–3.57 (m, 2 H), 2.38–2.34 (m, 1 H), 2.19–1.90 (m, 3 H + EtOAc); ^{13}C NMR (75 MHz, $CDCl_3$) δ (171.4, 171.1*), 154.3, (153.7, 153.6*), (148.1, 148.0*), (139.6, 139.5*), (128.1, 127.6*), (114.1, 114.0*), (110.8, 109.6*), 108.1, 102.0, (64.7, 64.3*), (58.8, 58.4), (56.7, 56.4), 49.0, (47.1, 46.3*), (30.9, 29.7*), (24.4, 23.5*).

dCA-NVOC-Proline

Prepared as above: FAB-MS $[M-H]^+$ 973.3, calcd for $C_{34}H_{42}N_{10}O_{20}P_2$ 972.2.

7.2.6 NVOC-Glutamine Cyanomethyl Ester



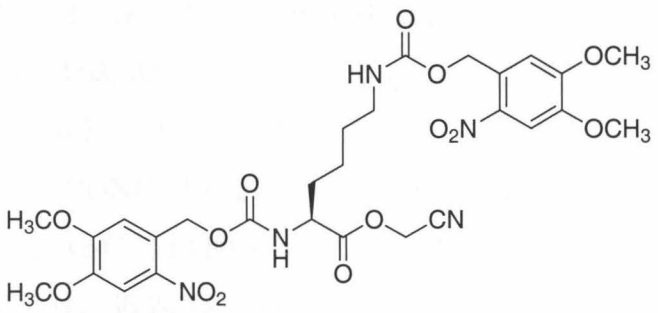
NVOC-Glutamine

The reaction was conducted as above. After addition of CH_2Cl_2 and 1 N $NaHSO_4$, a pale yellow precipitate formed. This was filtered to afford the product in 83% yield; FAB HRMS $[M-H]^+$ 386.1182, calcd for $C_{15}H_{19}N_3O_9$ 386.1200.

NVOC-Glutamine Cyanomethyl Ester

NVOC-Glutamine (50.8 mg, 0.13 mmol) was dissolved in 2 mL dry DMF. Triethylamine (55 μL, 0.39 mmol, 3 eq.) was added, followed by 2 mL chloroacetonitrile. After three hours the volatiles were removed in vacuo and the residue was purified by flash chromatography (silica, hexane→1:1 hexane/ethyl acetate→ethyl acetate) to afford 24 mg of an off-white solid (43% yield): ^{1}H NMR (300 MHz, DMSO-d_6) δ 8.15 (d, $J = 8$ Hz, 1 H, α-NH), 7.65 (s, 1 H, Ar-H), 7.18 (s, 1 H, Ar-H), 7.35 (s, 1 H, NH), 6.80 (s, 1 H, NH), 5.38 (m, 2 H, NVOC-CH_2), 5.00 (s, 2 H, OCH_2CN), 4.15 (m, 1 H, α-CH), 3.91 (s, 3 H, CH_3O), 3.86 (s, 3 H, CH_3O), 2.18 (m, 2 H, γ-CH_2), 1.97–1.79 (m, 2 H, δ-CH_2); ^{13}C NMR (75 MHz, DMSO-d_6) δ 173.6 COOR, 171.9 $CONH_2$, 156.2 CONH, 153.9 Ar-C, 148.2 Ar-C, 139.5 Ar-C, 128.2 Ar-C, 116.3 CN, 110.7 Ar-CH, 108.4 Ar-CH, 63.2 NVOC-CH_2, 56.7, 56.6 CH_3O's, 53.7 α-CH, 50.1 OCH_2CN, 31.3 γ-CH_2, 26.7 δ-CH_2; FAB HRMS M-Na^+ 447.1123, calcd for $C_{17}H_{20}N_4O_9Na$ 447.1128.

7.2.7 (NVOC)$_2$-Lysine Cyanomethyl Ester



(NVOC)$_2$-Lysine

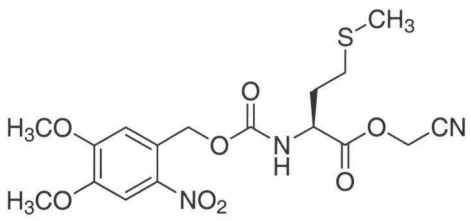
(NVOC)$_2$-Lysine was prepared as above using 250 mg (1.7 mmol) L-lysine, 943 mg (3.4 mmol, 2 eq.) NVOC-Cl and 360 mg (3.4 mmol, 2 eq.) sodium carbonate. Purified by flash chromatography (silica, 1:1 hexane/ethyl acetate→1% acetic acid in ethyl acetate) to give 830 mg of a pale orange solid (78% yield): ^{1}H NMR (300 MHz,

DMSO-d_6) δ 7.81 (d, $J = 8.1$ Hz, 1 H, α-NH), 7.67 (s, 1 H, Ar-H), 7.65 (s, 1 H, Ar-H), 7.43 (m, 1 H, ε-NH), 7.14 (s, 1 H, Ar-H), 7.15 (s, 1 H, Ar-H), 5.42 (m, 4 H, NVOC-CH_2's), 3.91 (s, 3 H, CH_3O), 3.89 (s, 3 H, CH_3O), 3.86 (s, 3 H, CH_3O), 3.85 (s, 3 H, CH_3O), 2.98 (m, 2 H, β-CH_2), 1.80–1.25 (m, 6 H, side chain CH_2's); ^{13}C NMR (75 MHz, DMSO-d_6) δ 173.7 COOH, 155.58, 155.67 CONH, 153.45, 153.26 Ar-C, 147.46, 147.59 Ar-C, 138.94, 139.27 Ar-C, 127.82, 128.11 Ar-C, 110.54, 109.99 Ar-CH, 108.11 (2) Ar-CH, 62.29, 62.12 NVOC-CH_2, 56.12, 56.02 CH_3O's, 53.74 α-CH, 39.5 (within solvent peak), 30.37, 28.82, 22.77 side chain CH_2's; FAB HRMS $[M\text{-}H]^+$ 625.1964, calcd for $C_{26}H_{32}N_4O_{14}$ 625.1993.

(NVOC)$_2$-Lysine Cyanomethyl Ester

(NVOC)$_2$-Lysine (103 mg, 0.16 mmol) was dissolved in dry DMF (1 mL). Triethylamine (67 μL, 0.48 mmol, 3 eq.) was added, followed by 1 mL chloroacetonitrile. After one hour the volatiles were removed in vacuo and the remaining yellow solid was purified by flash chromatography (silica, CH_2Cl_2 $\rightarrow$5:1 CH_2Cl_2/ethyl acetate$\rightarrow$ethyl acetate) to afford 106 mg of a pale yellow solid (100% yield): ^{1}H NMR (300 MHz, $CDCl_3$) δ 7.60 (s, 1 H, Ar-H), 7.58 (s, 1 H, Ar-H), 6.94 (2 s overlap, 2 H, Ar-H), 5.95 (s, 1 H, α-NH), 5.38 (m, 3 H, NVOC-CH_2 and ε-NH), 4.76 (m, 2 H, OCH_2CN), 4.35 (m, 1 H, α-CH), 3.93, 3.91, 3.88 (2 overlapping) (s, 3 H, CH_3O), 3.17 (m, 2 H, β-CH_2), 1.9–1.35 (m, 6 H, side chain CH_2's); ^{13}C NMR (75 MHz, $CDCl_3$) δ 170.89 COOR, 156.03, 156.2 CONH, 153.27, 153.45 Ar-C, 147.9 (2) Ar-C, 139.57, 139.41 Ar-C, 127.58, 127.42 Ar-C, 113.80 CN, 110.26, 109.92 Ar-CH, 107.96 (2) Ar-CH, 63.69, 63.31 NVOC-CH_2, 56.24 (2), 56.11 (2) CH_3O's, 53.41 α-CH, 48.74 OCH_2CN, 39.93, 30.77, 29.03, 21.91 side chain CH_2's; FAB HRMS $[M\text{-}H]^+$ 664.2079, calcd for $C_{28}H_{33}N_5O_{14}$ 664.2102.

7.2.8 NVOC-Methionine Cyanomethyl Ester



NVOC-Methionine

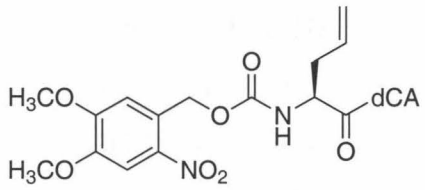
NVOC-Methionine was prepared in 98% yield, as above, starting with L-methionine (50 mg, 0.34 mmol): ^{1}H NMR (300 MHz, DMSO-d_6) δ 7.88 (d, J = 8.1 Hz, 1 H), 7.67 (s, 1 H), 7.14 (s, 1 H), 5.34 (dd, J = 15.0 Hz, 2 H), 4.18 (m, 1 H), 3.90 (s, 3 H), 3.84 (s, 3 H), 2.50 (m, 2 H), 2.05 (s, 3 H), 1.98 (m, 2 H); ^{13}C NMR (75 MHz, DMSO-d_6) δ 173.7, 155.9, 153.5, 147.7, 138.0, 128.3, 109.1, 62.5, 56.2, 56.1, 52.8, 30.3, 29.8, 14.5.

NVOC-Methionine Cyanomethyl Ester

NVOC-Methionine cyanomethyl ester was prepared in 82% yield, as above, starting with NVOC-L-methionine (32 mg, 0.08 mmol): ^{1}H NMR (300 MHz, $CDCl_3$) δ 7.75 (s, 1 H), 7.00 (s, 1 H), 5.65 (d, J = 15.0 Hz, 1 H), 5.53 (dd, J = 32.0 Hz, J = 15.1 Hz, 2 H), 4.80 (dd, J = 32.0 Hz, J = 15.1 Hz, 2 H), 4.63 (m, 1 H), 4.01 (s, 3 H), 3.96 (s, 3 H), 2.60 (m, 2 H), 2.30–2.10 (m, 2 H), 2.12 (s, 3 H).

7.3 Synthesis of Unnatural Amino Acids

7.3.1 dCA-NVOC-Allyl-Glycine



NVOC-Allyl-Glycine

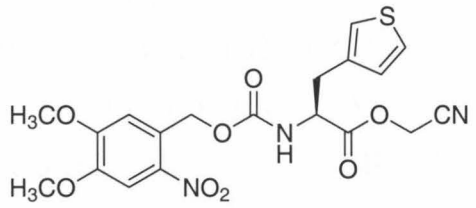
Allyl-glycine (100 mg, 0.87 mmol) and one equivalent of sodium carbonate (184 mg, 0.87 mmol) were dissolved in 15 mL water and 8 mL dioxane. A solution of one equivalent of NVOC-Cl (240 mg, 0.87 mmol) in 10 mL dioxane was added dropwise and the mixture was allowed to stir. After one hour, CH_2Cl_2 (25 mL) was added, followed by 1 N $NaHSO_4$ (25 mL). The aqueous layer was extracted three times with 25 mL CH_2Cl_2. The organic extracts were combined, dried over $MgSO_4$, filtered and concentrated in vacuo to give a pale yellow solid. The crude product was purified by flash chromatography (silica, 50:50 hexane/ethyl acetate $\rightarrow$1% acetic acid in ethyl acetate) to give 303 mg of a pale yellow solid (94% yield): ^{1}H NMR (300 MHz, DMSO-d_6) δ 7.89 (d, 1 H, J = 8.2 Hz, NH), 7.68 (s, 1 H, Ar-H), 7.15 (s, 1 H, Ar-H), 5.78 (dt, J = 16.8, 9.9, 6.8 Hz, 1 H, CH_2=CH), 5.38 (AB, J = 9.0 Hz, 1 H, NVOC-CH), 5.32 (AB, J = 9.0 Hz, 1 H, NVOC-CH), 5.13–5.02 (m, 2 H, CH_2=CH), 4.03 (m, 1 H, α-CH), 3.90 (s, 3 H, CH_3O), 3.84 (s, 3 H, CH_3O), 2.37 (m, 2 H, side chain CH_2); ^{13}C NMR (75 MHz, DMSO-d_6) δ 173.3 COOH, 155.8 CONH, 153.6 Ar-C, 147.7 Ar-C, 139.0 Ar-C, 134.3 CH_2=CH, 128.4 Ar-C, 117.8 CH_2=CH, 109.9 Ar-CH, 108.1 Ar-CH, 62.5 NVOC-CH_2, 56.3, 56.1 CH_3O's, 53.8 α-CH, 35.2 side chain CH_2.

NVOC-Allyl-Glycine Cyanomethyl Ester

NVOC-Allyl-glycine (100 mg, 0.27 mmol) was dissolved in 2 mL dry DMF. Triethylamine (113 μL, 0.81 mmol, 3 eq.) was added, followed by 2 mL chloroacetonitrile.

After 6 hours, the volatiles were removed in vacuo and the crude product was purified by flash chromatography (silica, hexane→8:2 hexane/ethyl acetate→1:1 hexane/ethyl acetate) to give 106 mg of a pale yellow oil (93% yield): ^{1}H NMR (300 MHz, $CDCl_3$) δ 7.20 (s, 1 H, Ar-H), 6.99 (s, 1 H, Ar- H), 5.71 (m, 1 H, CH_2=CH), 5.55 (*AB*, J = 15.0 Hz, 1 H, NVOC-CH), 5.47 (*AB*, J = 15.0 Hz, 1 H, NVOC-CH), 5.25–5.19 (m, 3 H, CH_2=CH and NH), 4.75 (dd, J = 20.7, 15.7 Hz, 2 H, OCH_2CN), 4.53 (m, 1 H, α-CH), 4.00 (s, 3 H, CH_3O), 3.95 (s, 3 H, CH_3O), 2.61 (m, 2 H, side chain CH_2); ^{13}C NMR (75 MHz, $CDCl_3$) δ 170.6 COOR, 155.4 CONH, 153.7 Ar-C, 148.1 Ar-C, 139.5 Ar-C, 131.5 CH_2=CH, 127.8 Ar-C, 120.4 CH_2=CH, 114.0 CN, 109.7 Ar-CH, 108.1 Ar-CH, 64.0 NVOC-CH_2, 56.5, 56.4 CH_3O's, 53.1 α-CH, 49.1 OCH_2CN, 36.0 side chain CH_2.

7.3.2 NVOC-β-(2-Thienyl)-L-Alanine Cyanomethyl Ester



NVOC-β-(2-Thienyl)-L-Alanine

NVOC-β-(2-Thienyl)-L-alanine was prepared as above using 35 mg (0.2 mmol) β-(2-Thienyl)-L-alanine, 56 mg (0.2 mmol, 1 eq.) NVOC-Cl and 43 mg (0.4 mmol, 2 eq.) sodium carbonate. Purified by flash chromatography (silica, hexane→7:3 hexane/ethyl acetate→1:1 hexane/ethyl acetate→1% acetic acid in ethyl acetate) to give 66 mg of pale yellow crystals (80% yield): ^{1}H NMR (300 MHz, DMSO-d_6) δ 7.98 (d, J = 8.4 Hz, 1 H, α-NH), 7.66 (s, 1 H, Ar-H), 7.31 (t, J = 3.1 Hz, 1 H, thiophene-H), 7.11 (s, 1 H, Ar-H), 6.92–6.90 (br d , 2 H, thiophene-H), 5.36 (*AB*, J = 13.0 Hz, 1 H, NVOC-CH), 5.30 (*AB*, J = 13.0 Hz, 1 H, NVOC-CH), 4.17 (m, 1

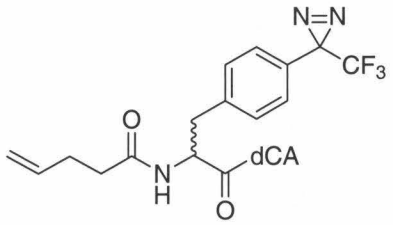
H, α-CH), 3.86 (s, 3 H, CH_3O), 3.82 (s, 3 H, CH_3O), 3.40–3.05 (m, 2 H, side chain CH_2); ^{13}C NMR (75 MHz, DMSO-d_6) δ 172.5 COOH, 155.6 CONH, 153.5 Ar-C, 147.7 Ar-C, 139.6 thiophene 2-C, 139.0 Ar-C, 128.0, 126.7, 126.1, 124.5 thiophene-C's and Ar-C, 110.0 Ar-CH, 108.2 Ar-CH, 62.3 NVOC-CH_2, 56.16, 56.06 CH_3O's, 55.5 α-CH, 30.68 side chain CH_2; HRMS CI$^+$ (NH_3) [M-NH_4]$^+$ 428.1145, calcd for $C_{17}H_{18}N_2O_8S$ + NH_3 428.1128.

NVOC-β-(2-Thienyl)-L-Alanine Cyanomethyl Ester

NVOC-β-(2-Thienyl)-L-alanine (60 mg, 0.15 mmol) was dissolved in 2 mL dry DMF. Triethylamine (61 μL, 0.44 mmol, 3 eq.) was added, followed by 2 mL chloroacetonitrile. After 9 hours, the volatiles were removed in vacuo and the crude product was purified by flash chromatography (silica, hexane→7:3 hexane/ethyl acetate→1:1 hexane/ethyl acetate) to give 61 mg of pale yellow crystals (93% yield): ^1H NMR (300 MHz, $CDCl_3$) δ 7.71 (s, 1 H, Ar-H), 7.21 (dd, J = 5.1, 0.75 Hz, 1 H, thiophene-H), 6.97 (overlapping s, 2 H, Ar-H, thiophene-H), 6.87 (d J = 3.5 Hz, 1 H, thiophene-H), 5.58 (AB, J = 6.5 Hz, 1 H, NVOC-CH), 5.52 (AB, J = 6.5 Hz, 1 H, NVOC-CH), 5.55 (m, 1 H, α-NH), 4.87–4.73 (m, 3 H, α-CH, CH_2CN), 3.97 (s, 3 H, CH_3O), 3.95 (s, 3 H, CH_3O), 3.42 (d, J = 6.5 Hz, 2 H, side chain CH_2); ^{13}C NMR (75 MHz, $CDCl_3$) δ 169.7 COOR, 155.1 CONH, 153.6 Ar-C, 148.0 Ar-C, 139.5 Ar-C, 136.0 thiophene C2, 127.6, 127.3, 127.2, 125.3 thiophene-C's and Ar-C, 113.6 CN, 109.7 Ar-CH, 108.0 Ar-CH, 64.0 NVOC-CH_2, 56.4, 56.3 CH_3O's, 54.5 α-CH, 49.1 CH_2CN, 31.8 side chain CH_2; HRMS CI$^+$ (NH_3) [M-NH_4]$^+$ 467.1225, calcd for $C_{18}H_{19}N_3O_8S$ + NH_3 467.1237.

7.4 Synthesis of Photoreactive Amino Acids

7.4.1 dCA-N-(4-PO)-3-[*p*-[3-(trifluoromethyl)-diazirin-3-yl]phenyl]-DL-alanine



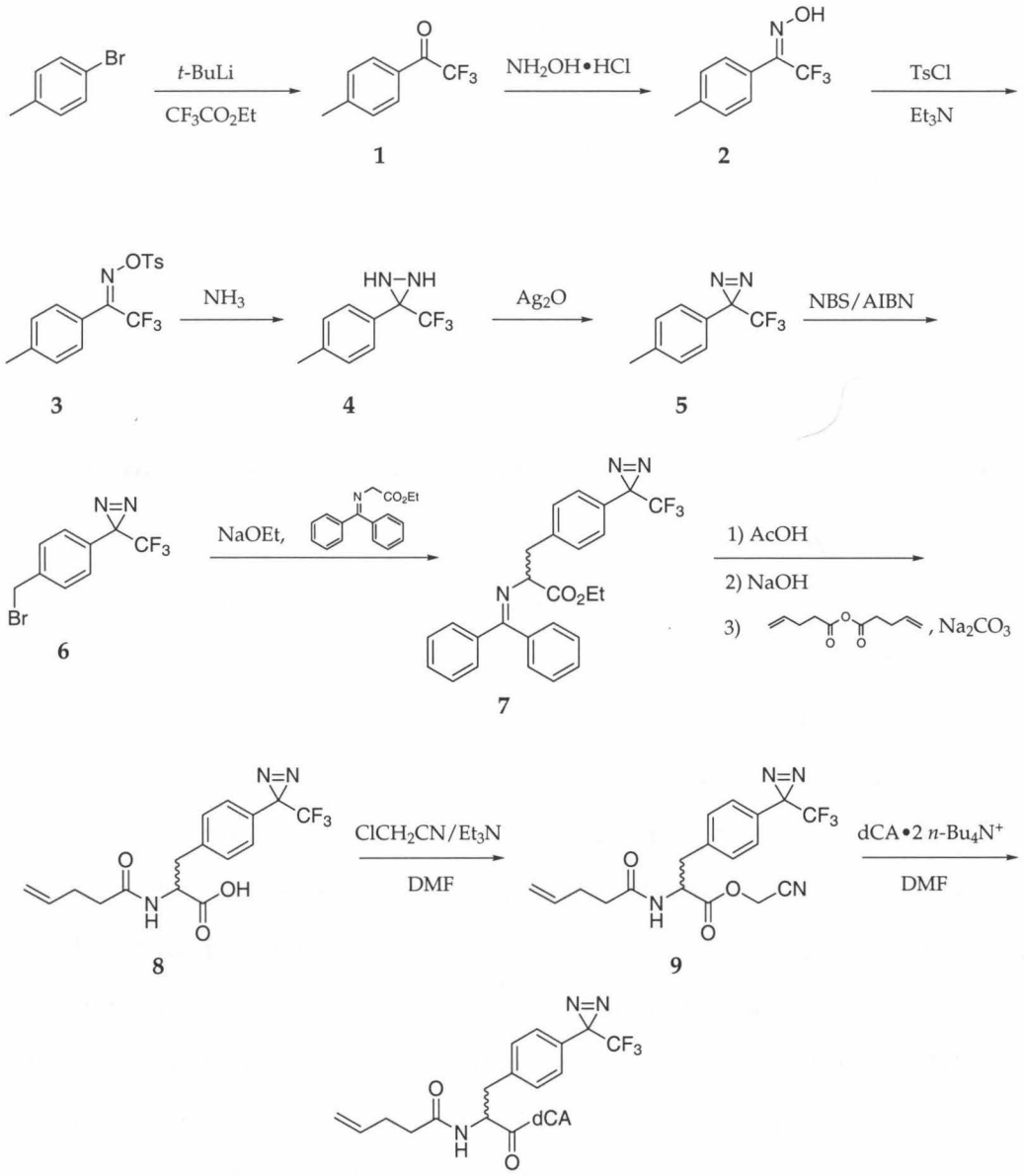
An overview of this synthesis is shown in Scheme 7.4. The various steps are adapted from several published procedures [8–10].

2,2,2-Trifluoro-1-(4-tolyl)ethanone (Compound 1)

Compound 1 was prepared with slight modification to the procedure of Platz *et al.* [10]. *t*-Butyllithium (100 mL, 1.7 M in pentane, 2.05 eq.) was cannulated into a dry three-neck 500 mL round-bottom flask at -78 °C. After 30 minutes, 4-bromotoluene (10.3 mL, 0.083 mol, 1.0 eq.) in 125 mL of dry THF was cannulated into the flask over a period of 15 minutes. After one hour, ethyl trifluoroacetate (19 mL, 0.16 mols, 1.9 eq.) was added dropwise in 0.5 mL portions. The mixture was allowed to warm in the CO_2/acetone bath for 14 hours, after which time it was quenched with water (25 mL), saturated ammonium chloride solution (15 mL) and an additional 100 mL water. The mixture was extracted four times with 125 mL portions of ether, the organics were combined, dried over $MgSO_4$, and concentrated in vacuo to give 20.5 grams of an orange liquid that was used without further purification. ^{1}H NMR of the crude product indicated a 54:46 ratio of desired product to starting material.

2,2,2-Trifluoro-1-(4-tolyl)ethanone Oxime (Compound 2)

To the crude mixture of 2,2,2-trifluoro-1-(4-tolyl)ethanone in a 500 mL round-bottom flask was added hydroxylamine·HCl (9.38 g, 0.14 mol, 3 eq. based on the amount



Scheme 7.4: Synthesis of dCA-N-(4-PO)-3-[*p*-[3-(trifluoromethyl)-diazirin-3-yl]phenyl]-DL-alanine.

of starting material determined by NMR) and NaOH (5.52 g, 137 mmol, 3 eq.) in absolute ethanol (250 mL). After refluxing for 28 hours, the mixture was concentrated to a volume of 50 mL. Water (100 mL) was added and the mixture was brought to pH 1 with 3 M HCl. The mixture was extracted four times with ether (100 mL portions), the organic layers were combined, dried over $MgSO_4$, and concentrated to give an orange oil. The crude product was purified by flash chromatography (silica,

CH_2Cl_2 eluent) to give white crystals conatminated with a small amount of an orange impurity that was removed by recrystallization from hexane to give a white solid (8.00 g) in 88% yield, 47% for 2 steps: ^{1}H NMR (300 MHz, $CDCl_3$) δ 9.85 (s, 1 H, NOH), 7.47 (d, J = 8.1 Hz, 2 H, Ar-H), 7.27 (d, J = 8.1 Hz, 2 H, Ar-H), 2.37 (s, 3 H,CH_3); ^{13}C NMR (75 MHz, $CDCl_3$) δ 147.64, 147.21 – 2 peaks for C=N 2 bond coupling to CF_3, 141.30 Ar-C1, 129.37, 128.70 Ar-C2,C6 and Ar-C3, C5, 122.83 Ar-C4, (q - 127, 122.57, 118.94, 115.5, J = 275 Hz, CF_3), 21.49 CH_3.

2,2,2-Trifluoro-1-(4-tolyl)ethanone O-(Toluenesulfonyl)oxime (Compound 3)

To a cooled solution (0 °C) of 2,2,2-trifluoro-1-(4-tolyl)ethanone oxime (7.85 g, 39 mmol, 1 eq.) in 150 mL toluene was added tosyl chloride (11.18 g, 58 mmol, 1.5 eq.) and triethylamine (8.7 mL, 1.6 eq.). After 15 minutes, the reaction was allowed to warm to room temperature. The volatiles were removed in vacuo, and the remaining residue was taken up in 100 mL ether. Water (100 mL) was added, and the mixture was extracted four times with 100 mL portions of ether. The organic layers were combined and washed with water (100 mL, followed by 50 mL), dried over $MgSO_4$, and concentrated to give an off white solid. The crude product was recrystallized from hexane to give a white solid (12.19 g, 88% yield): ^{1}H NMR (300 MHz, $CDCl_3$) δ 7.88 (d, J = 8.5 Hz, 2 H, Ts Ar-H), 7.38 (d, J = 8.3 Hz, 2 H, Ar-H), 7.32 (2 overlapping d, 4 H, Ar-H), 2.46 (s, 3 H,CH_3), 2.39 (s, 3 H,CH_3); ^{13}C NMR (75 MHz, $CDCl_3$) δ 154.0 (d, C=N 2 bond coupling to CF_3), 146.11 Ts-C1, 142.35 Ar-C1, 131.21, 129.87, 129.49, 129.26, 128.46, 121.61 - Ar, Ts-C's, (d - 121.55, 117.86, other two peaks buried, J = 278 Hz, CF_3), 21.78, 21.56 CH_3's.

3-(4-Tolyl)-3-(trifluoromethyl)diaziridine (Compound 4)

A 250 mL round-bottom flask charged with 2,2,2-trifluoro-1-(4-tolyl)ethanone O-(toluenesulfonyl)oxime was fitted with an ammonia condenser. The flask and the condenser were cooled with CO_2/acetone and anhydrous ammonia (150 mL) was condensed into the flask. The reaction was stirred at -78 °C for 9.5 hours, after which

time the ammonia was allowed to evaporate. The residue was dissolved in ether (100 mL) and 100 mL water was added. The mixture was extracted with ether (4x100 mL), the organic layers were combined, dried over $MgSO_4$, and concentrated to give 6.76 grams of a pale yellow solid that was used without further purification: ^{1}H NMR (300 MHz, $CDCl_3$) δ 7.47 (d, J = 8.0 Hz, 2 H, Ar-H), 7.20 (d, J = 8.0 Hz, 2 H, Ar-H), 2.77 (d, J = 8.6 Hz, 1 H), 2.36 (s, 3 H,CH_3), 2.22 (d, J = 8.6 Hz, 1 H). ^{13}C NMR (75 MHz, $CDCl_3$) δ 140.27 Ar-C4, 129.88 Ar-C1, 129.42, 128.04 Ar-C2,6 and 3,5, (d - 125.49, 121.80, other two peaks buried, J = 275 Hz, CF_3), 58.16, 57.69 q, J = 35 Hz C-$(NH)_2$, 21.28 CH_3.

3-(4-Tolyl)-3-(trifluoromethyl)diazirine (Compound 5)

To a solution of 3-(4-tolyl)-3-(trifluoromethyl)diaziridine (6.75 g, 33 mmol) in 100 mL dry ether was added 15 grams of freshly prepared Ag_2O. After 2 hours, another 22 grams of Ag_2O was added and the mixture was stirred for an additional two hours. The reaction mixture was filtered through Celite and washed with ether. The ether was removed using a rotary evaporator without heating to give an off-white solid which was purified by flash chromatography (silica, 9:1 hexane/ethyl acetate→ethyl acetate) to give 3.49 grams of a pale yellow liquid (51% yield, 59% based on recovered starting material): ^{1}H NMR (300 MHz, $CDCl_3$) δ 7.13 (d, J = 8.2 Hz, 2 H, Ar-H), 7.05 (d, J = 8.2 Hz, 2 H, Ar-H), 2.36 (s, 3 H, CH_3); ^{13}C NMR (75 MHz, $CDCl_3$) δ 140.02 Ar-C4, 129.70 Ar-C1, 126.56, 126.44 Ar-C2,6 and 3,5, 124.35, 120.71 - q, J = 274 Hz, CF_3, 28.86, 28.33 d, J = 40 Hz C<$(N)_2$, 22.92 CH_3.

3-(4-bromomethylphenyl)-3-(trifluoromethyl)diazirine (Compound 6)

To an argon purged solution of 3-(4-tolyl)-3-(trifluoromethyl)diazirine (3.33 g, 16.6 mmol) in 80 mL CCl_4 was added N-bromosuccinimide (2.95 g, 16.6 mmol). A reflux condenser was attached and the mixture was heated to 68 °C under an argon atmosphere. AIBN (100 mg, 0.6 mmol) was added and the mixture was allowed to reflux for 20 minutes. The mixture was cooled and the solid succinimide was filtered off. The filtrate was concentrated and purified by flash chromatography (silica, 20:1

petroleum ether:CH_2Cl_2) to give 2.40 grams of a colorless oil (52% yield): ^{1}H NMR (300 MHz, $CDCl_3$) δ 7.41 (d, J = 8.1 Hz, 2 H, Ar-H), 7.15 (d, J = 8.1 Hz, 2 H, Ar-H), 4.45 (s, 3 H,CH_2Br); ^{13}C NMR (75 MHz, $CDCl_3$) δ 139.64 Ar-C4, 129.58 Ar-C1, 127.17, 126.94 Ar-C2,6 and 3,5, 123.98, 120.35 - q, J = 274 Hz, CF_3, residual solvent obscures C<(N)$_2$, 29.17 CH_2Br.

N-(diphenylmethylene)-3-[*p*-[3-(trifluoromethyl)-diazirin-3-yl]phenyl]-DL-alanine ethyl ester) (Compound 7)

Sodium spheres (79 mg, 3.4 mmol) were placed in a dry 100 mL round-bottom flask under argon. Ethanol (7.8 mL) was added, followed by a solution of N-(diphenylmethylene)glycine ethyl ester (Aldrich) (880 mg, 3.3 mmol in 3.3 mL THF). The mixture was stirred for 45 min at °C, after which time, a solution of 3-(4-bromomethylphenyl)-3-(trifluoromethyl)diazirine (810 mg, 2.9 mmol in 5.5 mL dry THF) was added. After 4 hours, ether (100 mL) was added, and the organic layer was washed with sat. NH_4Cl (2 x 50 mL), and H_2O (50 mL), dried over $MgSO_4$, and concentrated to give a yellow oil. The product was purified by flash chromatography (silica, CH_2Cl_2→5% ether in CH_2Cl_2) to afford 485 mg of a yellow oil (36% yield): ^{1}H NMR (300 MHz, $CDCl_3$) δ 7.77 (d, J = 7.1 Hz, 1 H), 7.60 (d, J = 6.8 Hz, 1 H), 7.50 (m, 1 H), 7.40 (m, 1 H), 7.35–7.2 (m, 6 H), 7.05 (d, J = 8.2 Hz, 1 H), 7.00 (d, J = 8.2 Hz, 1 H), 4.25 (m, 1 H), 4.15 (m, 1 H), 3.35-3.15 (m, 2 H), 1.20 (t, J = 7.1 Hz, 3 H); ^{13}C NMR (75 MHz, $CDCl_3$) δ 171.3, 171.0, 140.4, 139.2, 135.9, 132.4, 130.5, 130.1, 128.8, 128.6, 128.3, 128.1, 127.6, 127.0, 126.4, 122.5 (q, J = 275 Hz), 67.0, 61.1, 39.3, 53.6, 28.4 (q, J = 40 Hz), 14.2.

N-(4-PO)-3-[*p*-[3-(trifluoromethyl)-diazirin-3-yl]phenyl]-DL-alanine) (Compound 8)

To a solution of **(7)** in 28 mL THF was added 7 mL 10 M AcOH. The solution was stirred for 24 hours, and another 7 mL 10 M AcOH was added. Volatiles were removed in vacuo, water (10 mL) was added, and the mixture was extracted with ether (2 x 25 mL). Conc. NaOH was added to the aqueous fraction to adjust the pH

to ≈13. This mixture was stirred at RT for 18 hours and neutralized with AcOH. The mixture was lyophilized, and the remaining solid was dissolved in a minimum amount of water. Dioxane (≈20 mL) was added, followed by a solution of 450 mg 4-PO-anhydride in 20 mL dioxane. After 7 hours, the reaction was purified using an ion-exchange column (Dowex 50x 4-400). The eluent was lyophilized and conc. HCl was added to acidify the mixture. The insoluble yellow solid was filtered and the remaining aqueous solution was extracted with CH_2Cl_2 (3 x 50 mL). The organic extracts were combined with the filtered solid, dried (Na_2SO_4), and concentrated to give a yellow oil. The product was purified by a flash column (silica, hexane→1:1 hexane/ethyl acetate→ethyl acetate→ ethyl acetate + 1% AcOH) to give 215 mg of a whitish solid (58% yield for 3 steps): ^{1}H NMR (300 MHz, $CDCl_3$) δ 11.01 (s - br), 7.22 (d, J = 8.3 Hz, 2 H), 7.10 (d, J = 8.0 Hz, 2 H), 6.95 (d, J = 7.8 Hz, 1 H), 5.71 (m, 1 H), 5.02–4.89 (m, 3 H), 4.11 (q, J = 7.2 Hz, 1 H), 3.24–3.18 (m, 2 H), 2.31 (m, 4 H); ^{13}C NMR (75 MHz, $CDCl_3$) δ 177.3, 174.2, 138.1, 136.5, 130.0, 128.0, 128.8 (q, J = 275 Hz), 126.7, 116.0, 53.2, 37.1, 35.4, 29.6, 28.5 (q, J = 40 Hz).

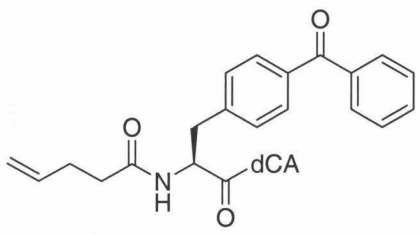
N-(4-PO)-3-[*p*-[3-(trifluoromethyl)-diazirin-3-yl]phenyl]-DL-alanine cyanomethyl ester) (Compound 9)

Prepared in 91% yield as above, starting with N-(4-PO-)-3-[*p*-[3-(trifluoromethyl)-diazirin-3-yl]phenyl]-DL-alanine: ^{1}H NMR (300 MHz, $CDCl_3$) δ 7.19 (d, J = 8.3 Hz, 2 H), 7.14 (d, J = 8.3 Hz, 2 H), 6.15 (d, J = 7.6 Hz, 1 H), 5.73 (m, 1 H), 5.10–4.90 (m, 2 H), 4.80 (dd, J = 15.7 Hz, 2 H), 3.18–3.12 (m, 2 H), 2.35–2.24 (m, 4 H); ^{13}C NMR (75 MHz, $CDCl_3$) δ 172.4, 170.2, 137.1, 136.5, 129.7, 128.4, 126.9, 122.0 (q, J = 275 Hz), 115.9, 113.8, 52.6, 49.1, 37.2, 35.3, 29.1, 29.0 (q, J = 40 Hz).

dCA-N-(4-PO-)-3-[*p*-[3-(trifluoromethyl)-diazirin-3-yl]phenyl]-DL-alanin

Prepared as above: FAB-MS $[M-H]^-$ 972.0, calcd for $C_{35}H_{39}N_{11}O_{15}P_2F_3$ 972.2.

7.4.2 dCA-4-PO-*p*-Benzoyl-Phenylalanine



4-PO-*p*-Benzoyl-Phenylalanine Cyanomethyl Ester

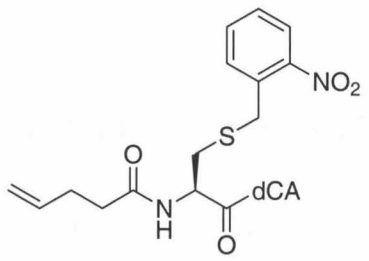
L-*p*-Benzoyl-phenylalanine·AcOH·H_2O (Bachem, 200 mg, 0.67 mmol) was mixed with Na_2CO_3 (284 mg, 2.68 mmol) in 30 mL 1:1 water/dioxane (not all solid dissolved). 4-PO-anhydride (125 mg, 0.67 mmol) in 5 mL dioxane was added and the mixture was stirred for 7 hours, after which 2 additional drops of 4-PO-anhydride were added. The mixture was stirred overnight, and CH_2Cl_2 (25 mL) and 1 N $NaHSO_4$ (25 mL) were added, and the mixture was extracted with CH_2Cl_2 (3 x 25 mL) and the combined organics were dried over Na_2SO_4. The volatiles were removed to give 600 mg of a pale yellow oil which was purified by flash chromatography (silica, ethyl acetate→ethyl acetate + 1% AcOH) to give 532 mg of a colorless oil. ^{1}H NMR indicated that there was still 4-PO-anhydride or acid present. 250 mg of the crude product was dissolved in DMF (2 mL) and chloroacetonitrile (2 mL) was added, followed by triethylamine (0.3 mL). After 39 hours (not optimized) the volatiles were removed and the compound was purified by flash chromatography (silica, 1:1 hexane/ethyl acetate) to give 131 mg of a clear oil (49% yield): ^{1}H NMR (300 MHz, $CDCl_3$) δ 7.76 (m, 4 H), 7.60 (t, J = 7.4 Hz, 1 H), 7.48 (t, J = 7.4 Hz, 2 H), 7.31 (d, J = 7.7 Hz, 2 H), 6.65 (d, J = 6.8 Hz, 1 H), 5.85–5.65 (m, 2 H), 5.05 (m, 2 H), 4.81 (dd, J = 15.0 Hz, J = 2.0 Hz, 2 H), 3.22 (*AB*, J = 7.2, 2 H), 2.32 (m - br, 4 H) (impurities: 4.08, 3.62); ^{13}C NMR (75 MHz, $CDCl_3$) δ 196.3, 172.5, 170.0, 140.3, 137.2, 136.4, 136.3, 132.4, 130.4, 129.8, 129.0, 128.2, 115.6, 113.8, 107.1, 90.6, 69.9, 65.8, 62.4, 52.5, 48.9, 37.1, 34.9, 29.0.

dCA-N-4-PO-*p*-Benzoyl-Phenylalanine

Prepared as above: FAB-MS $[M\text{-}H]^-$ 968.4, calcd for $C_{40}H_{45}N_9O_{16}P_2$ 968.8.

7.5 Synthesis of Caged Amino Acids

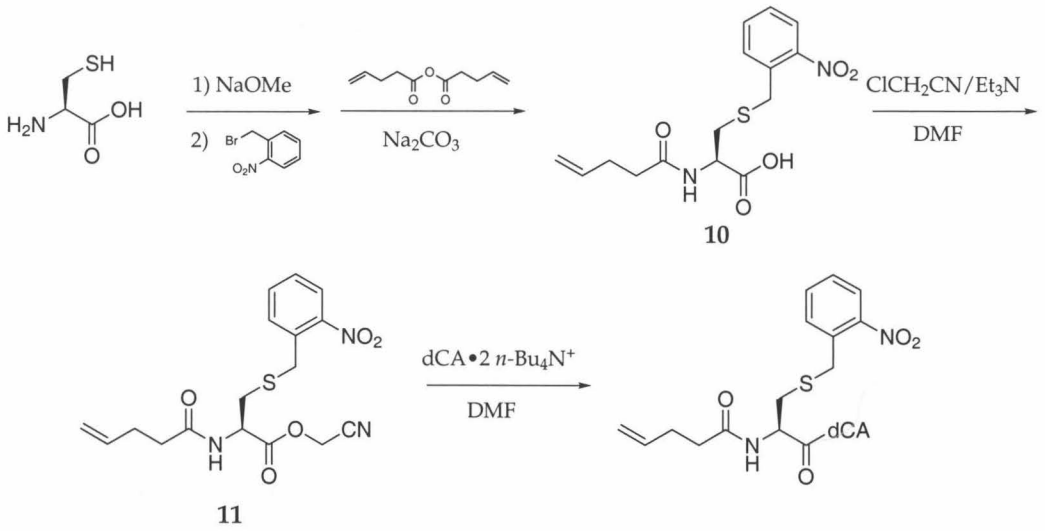
7.5.1 dCA-4-PO-Cysteine-S-*o*-Nitrobenzyl Thioether



An overview of this synthesis is shown in Scheme 7.5.

4-PO-Cysteine-S-*o*-Nitrobenzyl Thioether (Compound 10)

Sodium spheres (49 mg, 4.1 mmol) were dissolved in 4 mL dry MeOH at 0 °C and the mixture was warmed to room temperature over 10 min. L-Cysteine (250 mg, 2.06 mmol) was added, followed after 5 min. by *o*-nitrobenzyl bromide (356 mg, 1.65 mmol). After 1 hour, the mixture had gelled. Methanol (≈5 mL) was added, but the gel did not dissolve. Volatiles were removed on a rotovap, and the residue was dissolved in water (10 mL) and dioxane (10 mL). Na_2CO_3 was added (436 mg, 4.1 mmol), followed by 4-PO-anhydride (380 mg, 2.1 mmol). After 3 hours, an additional 50 mg 4-PO-anhydride was added. After an additional hour, CH_2Cl_2 (25 mL) and 1 N $NaHSO_4$ (25 mL) were added, and the mixture was extracted with CH_2Cl_2 (3 x 25 mL) and the combined organics were dried over Na_2SO_4. The volatiles were removed, and the product was purified by flash chromatography (silica, 1:1 hexane/ethyl acetate→ethyl acetate + 1% AcOH) to give 472 mg of a pale yellow solid (85% yield for 2 steps): ^{1}H NMR (300 MHz, CD_3OD) δ 7.93 (*AB*, J = 9.0 Hz, J = 1.3 Hz), 7.62–7.40 (m, 3 H), 5.90–5.75 (m, 1 H), 4.97 (dd, J = 32.0 Hz, J =



Scheme 7.5: Synthesis dCA-4-PO-Cysteine-S-*o*-Nitrobenzyl Thioether.

1.8 Hz, 2 H), 4.94 (dd, J = 32.0 Hz, J = 1.8 Hz, 2 H), 4.59 (dd, J = 4.8 Hz, 1 H), 4.08 (m, 2 H), 2.90–2.75 (m, 2 H), 2.34 (m, 4 H); ^{13}C NMR (75 MHz, CD_3OD) δ 175.4, 173.5, 149.7, 138.1, 134.9, 134.3, 133.3, 129.7, 126.5, 116.0, 53.0, 36.0, 34.4, 34.3, 30.7.

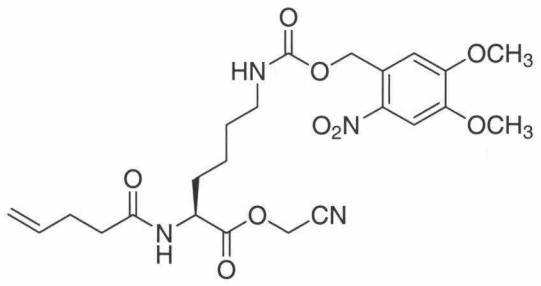
4-PO-Cysteine-S-*o*-Nitrobenzyl Thioether Cyanomethyl Ester (Compound 11)

Prepared in 85% yield, as above, starting with 4-PO-Cys-S-NB (360 mg, 1.06 mmol): ^{1}H NMR (300 MHz, $CDCl_3$) δ 8.00 (d, J = 8.2 Hz, 1 H), 7.60 (t, J = 7.7 Hz, 1 H), 7.44 (m, 2 H), 6.8 (m - br, 1 H), 5.83 (m, 1 H), 5.10–4.95 (m, 2 H), 4.80 (s, 2 H), 4.11 (m, 2 H), 2.94 (m, 2 H), 2.36 (m - br, 4 H); ^{13}C NMR (75 MHz, $CDCl_3$) δ 172.7, 169.5, 148.3, 136.7, 133.4, 133.2, 132.1, 128.7, 125.6, 115.7, 114.2, 51.6, 49.4, 35.0, 34.0, 33.4, 29.3.

dCA-4-PO-Cysteine-S-*o*-Nitrobenzyl Thioether (dCA-Caged-Cys)

Prepared as above: FAB-MS $[M-H]^+$ 957.2, calcd for $C_{34}H_{42}N_{10}O_{17}P_2S$ 956.2.

7.5.2 α-4-PO-ε-NVOC-Lysine Cyanomethyl Ester



α-4-PO-ε-BOC-Lysine

ε-BOC-Lysine (500 mg, 2.0 mmol) and Na_2CO_3 were stirred in 1:1 water/dioxane (30 mL). A solution of 4-PO-anhydride (407 mg, 2.23 mmol, 1.1 eq. in dioxane) was added and the mixture was stirred for 1 hour. CH_2Cl_2 (25 mL) and 1 N $NaHSO_4$ (25 mL) were added, and the mixture was extracted with CH_2Cl_2 (3 x 25 mL) and the combined organics were dried over Na_2SO_4. The volatiles were removed to give 1.27 g of a pale yellow oil. The product was purified by flash chromatography (silica, pet. ether→1:1 pet. ether/ethyl acetate→ethyl acetate→ethyl acetate + 1% AcOH) to give 600 mg of a colorless film (90% yield): ^{1}H NMR (300 MHz, CD_3OD) δ 5.82 (m, 1 H), 5.03 (d, J = 17.0 Hz, 1 H), 4.67 (d, J = 10.0 Hz, 1 H), 4.38 (AB, J =5.0 Hz, 1 H), 2.91 (t, J = 7.6 Hz, 2 H), 2.34 (m - br, 4 H), 1.95–1.38 (m, 6 H); ^{13}C NMR (75 MHz, CD_3OD) δ 175.6, 175.3, 138.2, 115.8, 53.3, 40.4, 36.0, 31.9, 30.8, 27.9, 23.8.

α-4-PO-Lysine

To a solution of α-4-PO-ε-BOC-lysine (600 mg, 1.83 mmol) and 1.25 g 1,4-dimethoxybenzene in 10 mL CH_2Cl_2 was added trifluoroacetic acid (10 mL). After 30 minutes, volatiles were removed (CO_2/acetone trap), toluene (15 mL) was added, and the volatiles were removed again. The remaining residue was triturated with ether (3 x 25 mL). Attempts to dissolve the mixture in MeOH and precipitate the product with ether failed, and the methanol/ether mixture was removed to yield 400 mg of an oil (96% yield) which was used directly in the synthesis of α-4-PO-ε-NVOC-lysine.

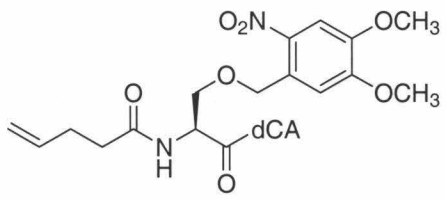
α-4-PO-ε-NVOC-Lysine

Prepared in 58% yield as above from 100 mg α-4-PO-lysine, 121 mg NVOC-Cl, and 93 mg Na_2CO_3: ^{1}H NMR (300 MHz, $CDCl_3$) δ 9.25 (s - br), 7.67 (s, 1 H), 7.01 (s, 1 H), 6.84 (d, J = 8.4 Hz, 1 H), 5.80 (m, 1 H), 5.60–5.40 (m, 3 H), 5.06 (m, 2 H), 4.56 (m, 1 H), 3.97 (s, 3 H), 3.94 (s, 3 H), 3.19 (d - br, 2 H), 2.35 (s - br, 4 H), 1.95-1.35 (m, 6 H); ^{13}C NMR (75 MHz, $CDCl_3$) δ 176.2, 175.4, 173.7, 156.5, 153.6, 148.1, 139.7, 136.7, 128.0, 115.7, 110.3, 108.1, 63.6, 56.4 (2), 52.2, 40.5, 35.4, 31.4, 29.5, 29.3, 22.3.

α-4-PO-ε-NVOC-Lysine Cyanomethyl Ester

Prepared in 81% yield as above, starting with 100 mg α-4-PO-ε-NVOC-lysine: ^{1}H NMR (300 MHz, $CDCl_3$) δ 7.69 (s, 1 H), 7.01 (s, 1 H), 6.52 (d - br, 1 H), 5.80 (m, 1 H), 5.48 (s, 2 H), 5.25 (t - br, 1 H), 5.01 (dd, J = 18.0 Hz, J = 1.2 Hz, 2 H), 4.88 (dd, J = 15.7 Hz, J = 2.0 Hz, 2 H), 4.61 (m, 1 H), 3.98 (s, 3 H), 3.96 (s, 3 H), 3.24 (m, 2 H), 2.35 (m, 4 H), 1.85 (m, 2 H), 1.58 (m, 2 H), 1.43 (m, 2 H); ^{13}C NMR (75 MHz, $CDCl_3$) δ 172.7, 171.1, 156.3, 153.6, 148.2, 139.8, 136.8, 127.9, 115.6, 114.2, 110.3, 108.2, 63.6, 56.5, 56.4, 51.8, 48.9, 40.2, 35.2, 30.9, 29.4, 29.3, 22.1.

7.5.3 dCA-4-PO-Serine-O-*o*-Nitroveratryl Ether



α-4-PO-Serine-O-*o*-Nitroveratryl Ether

Prepared in 82% yield as above, starting from the Na^+ salt of L-serine-O-NV [11]: ^{1}H NMR (300 MHz, $CDCl_3$) δ 7.62 (s, 1 H), 7.12 (s, 1 H), 6.79 (d, J = 7.1 Hz, 1 H), 5.80 (m, 1 H), 5.15–4.78 (m, 5 H), 4.10 (dd, J = 9.0 Hz, J = 3.2 Hz, 1 H), 3.82 (dd, J = 9.0 Hz, J = 3.2 Hz, 1 H), 3.94 (s, 3 H), 3.92 (s, 3 H), 2.39 (m, 4 H); ^{13}C NMR

(75 MHz, $CDCl_3$) δ 173.8, 173.5, 153.8, 147.7, 139.1, 136.5, 129.6, 115.9, 109.6, 107.9, 70.8, 70.3, 56.4, 56.3, 52.9, 35.4, 29.4.

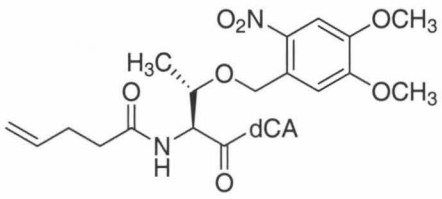
α-4-PO-Serine-O-o-Nitroveratryl Ether Cyanomethyl Ester

Prepared in 94% yield as above, starting from 90 mg α-4-PO-Ser-O-NV: ^{1}H NMR (300 MHz, $CDCl_3$) δ 7.63 (s, 1 H), 7.01 (s, 1 H), 6.52 (d, J = 7.1 Hz, 1 H), 5.85 (m, 1 H), 5.12–4.75 (m, 7 H), 4.11 (dd, J = 8.0 Hz, J = 3.2 Hz, 1 H), 3.83 (dd, J = 8.0 Hz, J = 3.2 Hz, 1 H), 4.00 (s, 3 H), 3.94 (s, 3 H), 2.40 (m, 4 H); ^{13}C NMR (75 MHz, $CDCl_3$) δ 172.4, 169.1, 153.7, 148.0, 139.5, 136.7, 128.7, 115.7, 113.9, 110.2, 108.0, 70.7, 70.6, 56.6, 56.4, 52.4, 49.2, 35.3, 29.3.

dCA-4-PO-Serine-O-o-Nitroveratryl Ether

Prepared as above: FAB-MS $[M-H]^-$ 999.3, calcd for $C_{36}H_{46}N_{10}O_{20}P_2$ 1000.24.

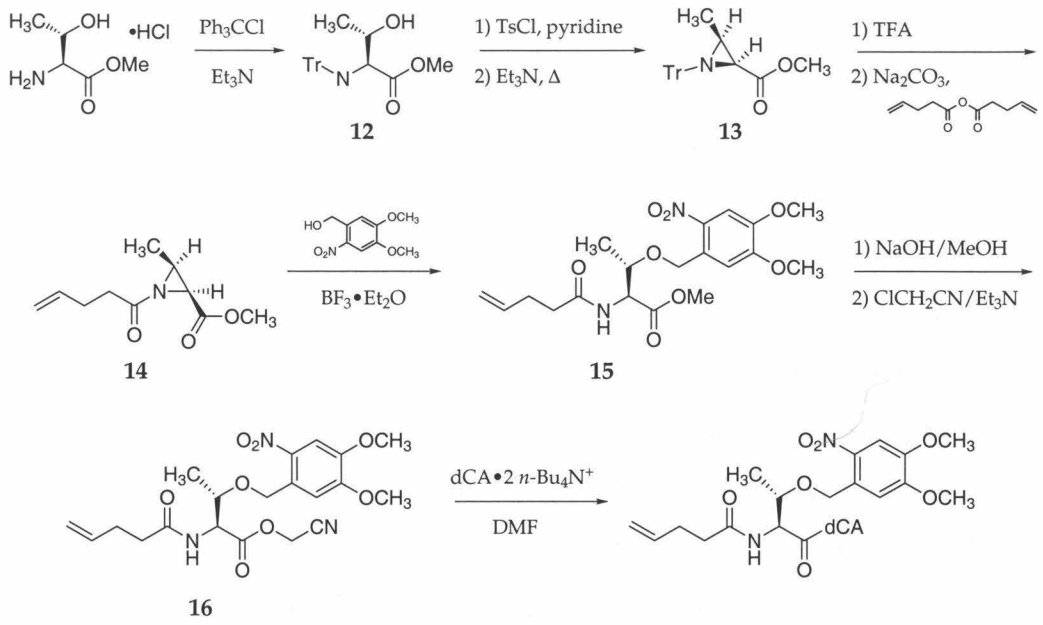
7.5.4 dCA-4-PO-Threonine-O-o-Nitroveratryl Ether



An overview of this synthesis is shown in Scheme 7.6. The procedure is adapted from [12].

α-N-Trityl-Threonine Methyl Ester (Compound 12)

L-Threonine methyl ester·HCl (1.0 g, 5.9 mmol, 1.0 eq) was dissolved in 50 mL dry CH_2Cl_2 in a dry 100 mL round-bottom flask under argon at 0 °C. Triethylamine (1.19 g, 11.8 mmol, 2.0 eq) was added, followed by trityl chloride (1.64 g, 5.9 mmol, 1.0 eq). After 10 minutes at 0 °C, the mixture was allowed to warm to room temperature and was stirred for an additional hour. A solution of 10% citric acid (20 mL) was added,



Scheme 7.6: Synthesis dCA-4-PO-Threonine-O-*o*-Nitroveratryl Ether.

and the organic layer was extracted, washed with water, and dried over Na_2SO_4 with activated charcoal. The mixture was filtered and concentrated in vacuo to give 1.82 g of a foamy white solid (83% yield): ^{1}H NMR (300 MHz, $CDCl_3$) δ 7.48 (d, J = 7.4 Hz, 6 H), 7.30–7.10 (m, 9 H), 3.81 (m, 1 H), 3.39 (d, J = 7.3 Hz, 1 H), 1.19 (d, J = 7.3 Hz, 3 H); ^{13}C NMR (75 MHz, $CDCl_3$) δ 173.6, 145.4, 128.9, 127.8, 126.6, 70.7, 69.7, 62.4, 51.6, 18.9.

Compound 13

α-N-trityl-threonine methyl ester (1.74 g, 4.63 mmol) was dissolved in 10 mL dry pyridine in a dry 100 mL round-bottom flask under argon. In a separate dry 100 mL round-bottom flask under argon, tosyl chloride (2.65 g, 13.9 mmol, 3.0 eq) was dissolved in 10 mL dry pyridine. The amino acid solution was cooled to 0 °C, and the TsCl solution was cannulated in. The reaction was allowed to warm to room temperature, and was stirred for 20 hours. The pyridine was removed in vacuo, and EtOAc (25 mL) and water (25 mL) were added and the mixture was extracted. The aqueous layer was further extracted with EtOAc (2 x 25 mL), and the combined organic extracts were washed with 10% citric acid (3 x 10 mL) and water (2 x 10 mL).

The organic extracts were dried over Na_2SO_4 and concentrated in vacuo to give 2.66 g of an orange oil. 2.54 g of the oil was dissolved in dry THF (20 mL) under argon, triethylamine (937 mg, 2.0 eq) was added, and the mixture was refluxed overnight. The solvent was removed in vacuo, and the residue was dissolved in EtOAc and 10% citric acid and extracted with EtOAc (3 x 25 mL). The combined organic extracts were washed with 1 M $NaHCO_3$ and water, dried over Na_2SO_4 and concentrated in vacuo to give 1.66 g of a dark brown solid. The product was purified by flash chromatography (silica, 4:1 hexane/CH_2Cl_2→1:1 hexane/CH_2Cl_2→1:3 hexane/CH_2Cl_2 to give 804 mg of a foamy white solid (yield: 49% for 2 steps. Repeated on larger scale, yield: 78%): ^{1}H NMR (300 MHz, $CDCl_3$) δ 7.52 (d, J = 7.2 Hz, 6 H), 7.27–7.15 (m, 9 H), 3.69 (s, 3 H), 1.89 (d, J = 6.4 Hz, 1 H), 1.63 (m, 1 H), 1.36 (d, J = 5.5 Hz, 3 H); ^{13}C NMR (75 MHz, $CDCl_3$) δ 170.6, 143.9, 129.3, 127.6, 126.8, 75.0, 51.7, 35.9, 34.8, 13.3.

Compund 14

To a solution of **13** (400 mg, 1.1 mmol) in 5 mL 1:1 MeOH/CH_2Cl_2 at 0 °C was added TFA (5 mL). The reaction was stirred at 0 °C for 10 min and allowed to warm to RT. After 3 hours, an additional 5 mL TFA was added and the mixture was stirred overnight. The solvent was removed by forming an azeotrope with toluene. Water (20 mL) and dioxane (20 mL) were added, along with 4-PO-anhydride (261 mg, 1.43 mmol, 1.3 eq) and Na_2CO_3 (466 mg, 4 eq). After 5 hours, CH_2Cl_2 (25 mL) was added, and the mixture was extracted with CH_2Cl_2 (3 x 25 mL). The combined organic extracts were dried over Na_2SO_4 and concentrated in vacuo to give 660 mg of a colorless oil. The product was purified by flash chromatography (silica, 9:1 hexane/EtOAc→1:1 hexane/EtOAc→EtOAc to give 184 mg of a colorless oil (yield: 84% for 2 steps): ^{1}H NMR (300 MHz, $CDCl_3$) δ 5.82 (m, 1 H), 5.10–4.97 (m, 2 H), 3.79 (s, 3 H), 3.20 (d, J = 6.4 Hz, 1 H), 2.81 (m, 1 H), 2.53 (m, 2H), 2.39 (m, 2 H), 1.37 (d, J = 5.6 Hz, 3 H); ^{13}C NMR (75 MHz, $CDCl_3$) δ 183.5, 167.8, 136.7, 115.7, 52.4, 38.7, 37.9, 35.6, 28.6, 13.1.

α-4-PO-Threonine-O-o-Nitroveratryl Ether Cyanomethyl Ester (Compound 16)

Compound **14** (170 mg, 0.86 mmol, 1.0 eq) and o-nitroveratryl alcohol (NVOH, 550 mg, 2.58 mmol, 3.0 eq) were placed in dry 100 mL round-bottom flask under argon. Dry $CDCl_3$ (30 mL) was added ($CDCl_3$ was used because it is not stabilized with EtOH). Most but not all of the NVOH dissolved and the flask was cooled to 0 °C. 5 drops of $BF_3{\cdot}OEt_2$ were added and after 15 min, the mixture was warmed to RT and allowed to stir overnight, after which time, an additional 100 μL $BF_3{\cdot}OEt_2$ was added. After 6 additional hours, 5% $NaHCO_3$ was added, and the mixture was extracted with CH_2Cl_2 (3 x 25 mL). The combined organic extracts were dried over Na_2SO_4 and concentrated in vacuo to give 652 mg of a yellow solid. Attempts to separate **15** from NVOH by TLC failed, and the decision was made to hydrolyze the ester directly. The solid was dissolved in MeOH (30 mL), and 1 N NaOH (15 mL) was added. After 6 hours, additional 1 N NaOH (5 mL) was added and the reaction was stirred overnight. The MeOH was removed by rotary evaporation, and the mixture was extracted with CH_2Cl_2 (1 x50 mL, 2 x 25 mL) to remove NVOH. CH_2Cl_2 (50 mL) and 1 N $NaHSO_4$ (50 mL) were added to the aqueous layer, and it was extracted with CH_2Cl_2 (3 x 25 mL). The combined organic extracts were dried over Na_2SO_4 and concentrated in vacuo to give 107 mg of an orange-brown oil. The product was purified by flash chromatography (silica, 1:1 hexane/ethyl acetate→ethyl acetate→ethyl acetate + 1% AcOH). The product-containing fractions were decolorized with activated charcoal and concentrated to give 95 mg of a pale yellow oil. ^{1}H NMR indicated that the product was still not pure, and it was decided to continue to the cyanomethylation. 58 mg of the crude product and 50 μL of Et_3N were dissolved in 2 mL dry DMF. Chloroacetonitrile (2 mL) was added and the mixture was stirred overnight. Volatiles were removed in vacuo and the product was purified by flash chromatography (silica, 1:1 hexane/ethyl acetate) to give 49 mg of a pale yellow oil. (Yield: 77% for final step, 21% for 3 steps): ^{1}H NMR (300 MHz, $CDCl_3$) δ 7.63 (s, 1 H), 7.05 (s, 1 H), 6.29 (d, J = 10.1 Hz, 1 H), 5.85 (m, 1 H),

5.17–4.65 (m, 6 H), 4.30 (m, 1 H), 4.02 (s, 3 H), 3.94 (s, 3 H), 2.42 (m, 4 H), 1.34 (d, $J = 6.2$ Hz, 3 H); ^{13}C NMR (75 MHz, $CDCl_3$) δ 172.9, 169.6, 153.6, 148.0, 139.6, 136.7, 128.7, 115.8, 114.0, 110.8, 108.0, 75.6, 68.1, 56.6, 56.5, 56.4, 49.1, 35.5, 29.4, 16.0.

dCA-4-PO-Threonine-O-*o*-Nitroveratryl Ether

Prepared as above: FAB-MS $[M\text{-}H]^-$ 1013.3 (also see 1254.7), calcd for $C_{37}H_{48}N_{10}O_{20}P_2$ 1014.25.

7.6 Molecular Biology

7.6.1 General Procedure for Site-Directed Mutagenesis

In most instances, site-directed mutants were prepared using a standard two-step PCR-based cassette mutagenesis procedure using *Pfu* Polymerase (Stratagene). For each mutant, an enzymatically trimmed PCR-generated fragment was subcloned into the dephosphorylated fragment of the desired vector. In some cases, it was possible to synthesize the inserted mutant fragment without resorting to the PCR method. When possible, silent mutations were introduced to facilitate screening of positive colonies by restriction endonuclease digestion. We have found the *Bfa*I recognition sequence CTAG to be particularly helpful for screening TAG mutants. All mutants were verified by automated sequencing of the entire PCR cassette across both ligation sites.

7.6.2 General Procedure for In Vitro Transcription of mRNA

mRNA was synthesized in vitro from linearized plasmid DNA (1 μg/40 μL reaction) using the T7 MessageMachine kit from Ambion per the manufacturer's instructions. The products were purified by phenol:chloroform:isoamyl alcohol extraction and isopropanol precipitation. Purity of the mRNA was assayed by gel electrophoresis.

7.6.3 Mutants for Nonsense Suppression in the nAChR

The following is a list of mutants prepared for nonsense codon suppression in the nAChR from mouse muscle. All mutants are in the pAMV-PA vector [13].

αW67→TAG
αN68→TAG
αP69→TAG
αD70→TAG
αD71→TAG
αY72→TAG
αG73→TAG
αG74→TAG
αV75→TAG
αK76→TAG
αD89→TAG
αF124→TAG
γY117→TAG

Table 7.1: TAG mutants in the nAChR

In addition to these mutants, two other TAG mutants were prepared in other constructs. The first (α210GPI D70→TAG) was prepared in a truncated N-terminal fragment of the mouse nAChR α subunit fused with a sequence directing membrane anchorage via a glycosyl-phosphatidylinositol (GPI) linkage [14]. The second is the αW54→TAG mutant, prepared in the α7 neuronal nicotinic subunit from the mouse. This residue is homologous to the $\gamma 55/\delta 57$ mutants discussed in Chapter 2.

7.6.4 Conventional Mutants in the nAChR

In addition to the mutants outlined above, a series of mutants was prepared to investigate the roles of putative glycosylation sites in the nAChR. Sequence alignnments of the ligand-gated ion channel (LGIC) superfamily reveal an interesting trend. In the main immunogenic region of the α subunit of the nAChR, there is a proline residue at position 69. In the aligned positions of the LGICs, this proline is often replaced with a consensus asparagine-linked glycosylation sequence (NX($\frac{\mathrm{S}}{\mathrm{T}}$)). Both prolines and N-linked glycosylation are thought to induce β turns in protein structures [15, 16].

αP69→A	αP69→A, D70→S
βD68→N, P69→I	βD68→N, P69→I, A70→S
γD68→N, P69→I	γD68→N, P69→I, A70→S
δT72→D	δI71→P, T72→D
α7S69→D	α7V68→P, S69→D

Table 7.2: Conventional site-directed mutants to investigate the role of prolines and glycosylation sites in the extracellular domain.

Thus, several mutants were made in various subunits of both the muscle and α7 neuronal nAChR from the mouse to explore whether either a proline or a glycosylation site is vital to the function of the receptor, and if so, to determine if prolines and glycosylation sites are interchangable. The mutants are shown in Table 7.2. Although these mutants were not tested extensively, early experiments comparing the αP69→A mutant (which lacks the proline and cannot be glycosylated) and the αP69→A, D70→S mutant (which lacks the proline, but may be glycosylated) suggest that the αP69→A, D70→S expresses better (as evidenced by larger whole-cell currents in response to ACh) than the αP69→A mutant (Figure 7.1). This may imply that a turn-promoting unit in this region of the receptor is necessary for receptor function. This is consistent with the NMR structure of the MIR peptide shown in Chapter 6.

7.7 Oocyte Experiments

7.7.1 Oocyte Preparation and Injection

Oocytes were removed from *Xenopus laevis* as described previously [17] and maintained at 18 °C in ND96 solution consisting of 96 mM NaCl, 2 mM KCl, 1 mM $MgCl_2$, 1.8 mM $CaCl_2$, 5 mM HEPES (pH 7.5), supplemented with sodium pyruvate (2.5 mM), gentamicin (50 μg/mL), theophylline (0.6 mM) and horse serum (5%). Prior to microinjection, the NVOC-aminoacyl-tRNA (1 μg/μL in 1 mM NaOAc, pH 5.0) was deprotected by irradiating the sample for five minutes with a 1000 W xenon arc lamp (Oriel) operating at 600 W equipped with WG-335 and UG-11 filters (Schott).

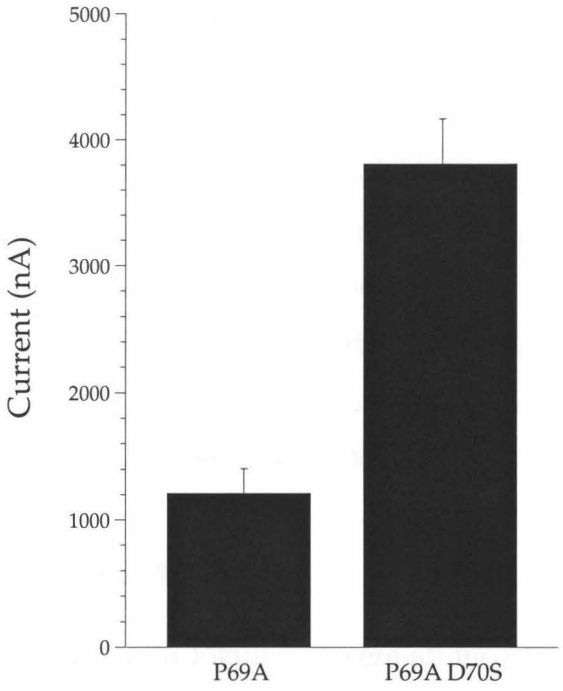


Figure 7.1: Whole-cell currents in response to 10 μM ACh for oocytes expressing nAChR αP69→A or αP69→A, D70→S.

4-PO-protected tRNAs were deprotected by treatment with 1 volume of a saturated solution of I_2 in 1 mM NaOAc, pH 5.0. The deprotected aminoacyl-tRNA was mixed 1:1 with the desired mRNA. In general, for experiments with suppression in the α subunit of the nAChR, oocytes were injected with 50 nL of a mixture containing 25 ng aminoacyl-tRNA and 1.2 ng of total mRNA (4:1:1:1 α:β:γ:δ subunit stoichiometry).

7.7.2 Electrophysiology

Two-electrode voltage clamp measurements were performed 36–48 hours after injection using a GeneClamp 500 amplifier (Axon Instruments, Foster City, CA). Microelectrodes were filled with 3M KCl and had resistances of 0.5 to 2 MΩ. Oocytes were perfused with a nominally calcium-free bath solution containing 96 mM NaCl, 2 mM KCl, 1 mM $MgCl_2$, 5 mM HEPES (pH 7.5). Macroscopic ACh-induced currents were recorded in response to application of 200 μM ACh at a holding potential of -80 mV. Whole-cell current measurements are reported as the mean $\pm$ s.e.m.

7.8 Results of Nonsense Suppression Experiments

In this section, we outline the results of nonsense suppression experiments in the nicotinic acetylcholine receptor. For each experiment, a successful incorporation is one in which the whole-cell currents in response to 200 μM ACh for the mutant receptor are greater than than those obtained when an identical mRNA mixture and an uncharged tRNA (tRNA–dCA) are injected. In most cases, this "dCA control" produced very small currents, however, at αN68, and several sites in the γ subunit, the dCA currents were not negligible (>500–1000 nA).

7.8.1 Natural Amino Acids

The natural amino acids shown in Figure 7.2 were all incorporated into various sites within the nAChR. Phenylalanine (protected with both the NPS and 4-PO groups) was successfully incorporated at position αE180 in the nAChR, thus demonstrating the viability of these protecting groups. Valine was successfully incorporated at αG73 and αV75, but no currents were seen at αG74. Tryptophan was successfully incorporated at two sites, αW67 and αY72. Proline was incorporated at a natural proline site, αP69 – no other sites were tested. Tyrosine was incorporated at several sites shown in Table 7.3 and was often used as a positive control for expression when the natural amino acid was not available.

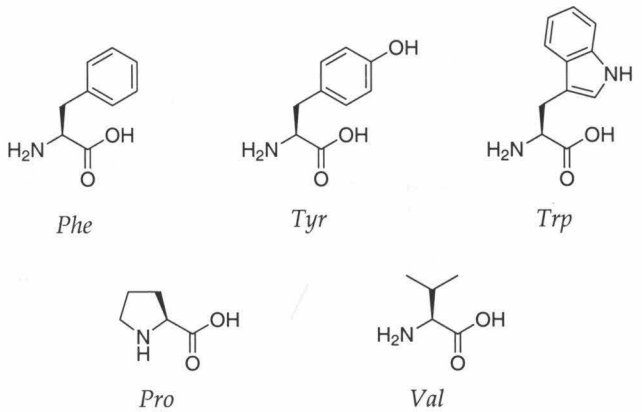


Figure 7.2: Natural amino acids incorporated into the nAChR by nonsense suppression.

Successes	αN68	αD70	αD71
	αY72	αK76	
Failures	αW67	αP69	

Table 7.3: Incorporation of tyrosine into the nAChR.

7.8.2 Biotinylated Amino Acids

The biotinylated amino acids shown in Figure 7.3 were used in several suppression experiments. In general, biocytin[2] was efficiently incorporated into the nAChR (Table 7.4), with whole-cell currents that were comparable to natural amino acids at the same position (>5 μA). The mid-chain biocytin analog could also be incorporated at several sites in the receptor (Table 7.5). Although the currents were often well above background (500–1000 nA), they were not large enough to allow biochemical experiments (such as the binding of streptavidin) to be performed. Finally, we were unable to see expression of the long-chain biocytin analog at any of the sites tested (αD70, αK76, and αE180).

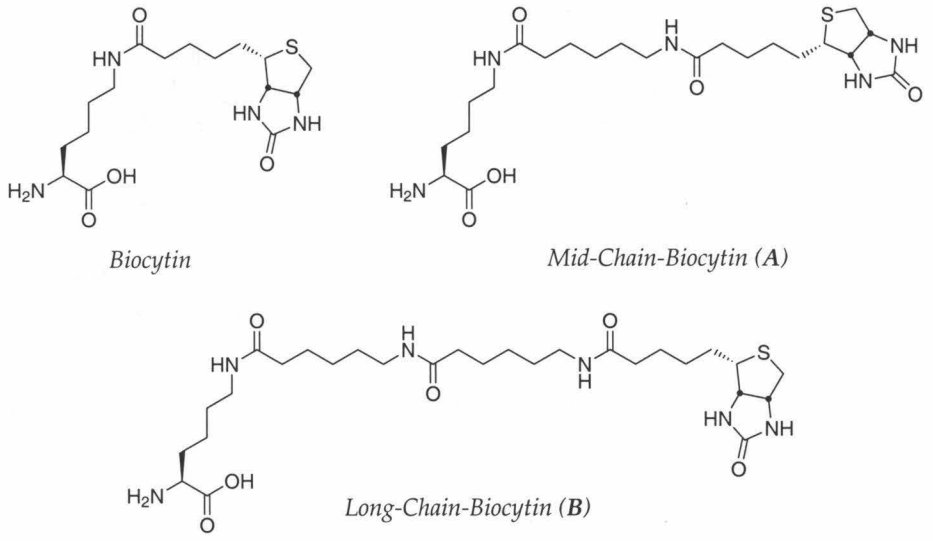


Figure 7.3: Biotinylated amino acids incorporated into the nAChR by nonsense suppression

[2]For full details, see Chapter 6

Successes	αN68	αD70	αD71	αV75
	αK76	αE180	γS129	γS131
	γV132	γT133	γY134	γF135
	γF137			
Failures	αW67	αP69	αY72	αG73
	αG74			

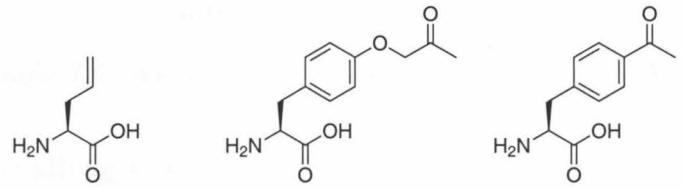
Table 7.4: Biocytin incorporation into the nAChR.

Successes	αN68	αD70	αD71
Failures	αP69		

Table 7.5: Mid-chain-biocytin incorporation into the nAChR.

7.8.3 Chemically Reactive Amino Acids

In this section, we outline the incorporation of amino acids with chemically reactive side chains. Such amino acids might be useful for post-translational modification of proteins and may complement methods for site-specific cell labeling, such as SCAM and the biotinylated amino acids described above.



Allyl-Gly *Keto-Tyr* *Acetyl-Phe*

Figure 7.4: Amino acids with chemically reactive side chains incorporated into the nAChR by nonsense suppression.

The first amino acid is allyl-glycine, which contains a terminal olefin. Since terminal olefins are not common on the surfaces of cells, we reasoned that this amino acid might display unique reactivity. Water-soluble olefin metathesis catalysts, such as the one shown in Figure 7.5 can mediate the reaction shown in Scheme 7.7 [18, 19]. By using an appropriately substituted catalyst, we can efficiently and selectively transfer a specific functionality to a protein containing a terminal olefin. To this end, we incorporated allyl-glycine at several sites in the nAChR. Whole-cell currents were uniformally large (often greater than 10 μA) at several sites shown in Table 7.6. Unfortunately, at high concentrations, the catalyst shown in Figure 7.5 is acutely toxic

to cells (Figure 7.6), and this approach may be limited.

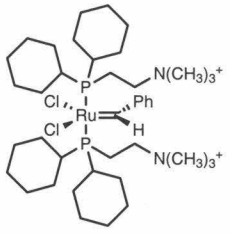
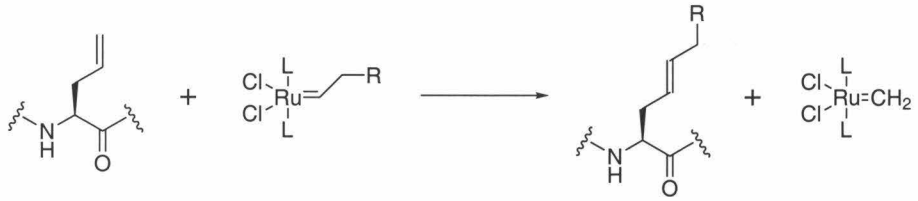


Figure 7.5: A water-soluble olefin metathesis catalyst.



Scheme 7.7: Olefin metathesis

Successes	αD70	αK76	αE180
Failures	None		

Table 7.6: Allyl-glycine incorporation into the nAChR.

In addition to allyl-glycine, two other amino acids with specific side chain reactivies were incorporated into the nAChR at position αD70. Both keto-tyrosine [20, 21] and 4-acetyl-phenylalanine [21] gave whole-cell currents in excess of 1 μA in response to 200 μM ACh. Since ketones are not prevalent on the cell surface, it may be possible to label these amino acids with functionalized hydrazines or hydroxylamines [20, 22].

7.8.4 Photochemically Reactive Amino Acids

Finally, we attempted to incorporate several amino acids with photo-reactive side chains into the nAChR (Figure 7.7). Benzoyl-phenylalanine contains a benzophenone moiety, which upon irradiation with UV light, forms a triplet excited state that can efficiently abstract hydrogen atoms from a variety of C–H bonds. The remaining carbon-centered radicals may recombine to form new carbon-carbon bonds. Thus, this amino acid may be used to cross-link separate subunits in a multi-subunit protein

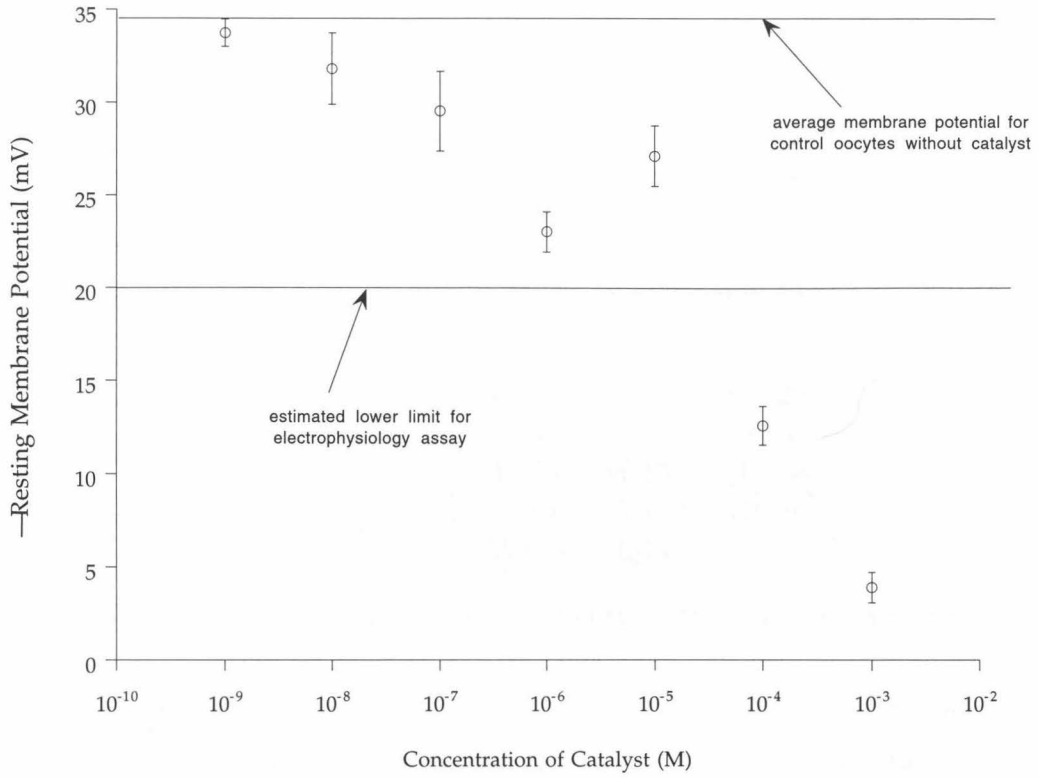
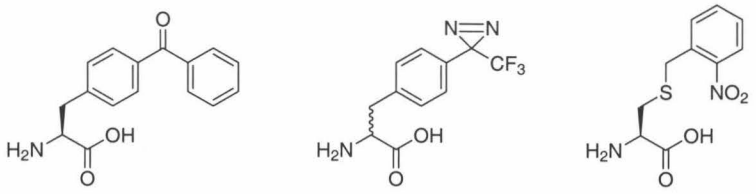


Figure 7.6: Resting membrane potentials for oocytes treated with the catalyst shown in Figure 7.5. Each point represents the average resting membrane potential of 5 oocytes treated for 15 minutes with the given concentration of catalyst in degassed ND96, pH 6.0. Catalyst concentrations >10 μM rapidly kill the cells.

such as the nAChR, and allow us to determine which amino acids lie at subunit interfaces. The results of attempts to incorporate benzoyl-phenylalanine at several sites in the nAChR are shown in Table 7.7. Benzoyl-phe was successfully incorporated at several sites, including positions where the wild-type residue is either a polar or charged amino acid. However, in many cases, especially in the γ subunit, where dCA currents were high, the whole-cell currents were not much larger than the dCA control. Nonetheless, this amino acid may be useful for photocross-linking studies when the appropriate biochemical assays to identify successful cross-linking (such as gel-shift assays, or mass-spectrometry) are perfected.

In addition to benzoyl-phenylalanine, we also attempted to incoporate a diazirine-containing amino acid (Figure 7.7) into the nAChR. Although this amino acid has been incorporated into proteins via nonsense suppression [23], we did not observe



Benzoyl-Phe *Diazirine* *Caged-Cys*

Figure 7.7: Amino acids with photochemically reactive side chains incorporated into the nAChR by nonsense suppression.

Successes	αD70	αY72	γS129
	γS131	γT133	γY134
	γF135	γF137	γD138
Failures	γI130	γV132	γP136
	γW139	γQ140	

Table 7.7: Benzoyl-phenylalanine incorporation into the nAChR.

ACh-induced currents during attempts to incorporate this amino acid at α70. Although α70 is a permissive site, it is possible that this residue may be incompatible with α70, and further testing at other sites may be fruitful. Fortunately, benzoyl-phenylalanine is tolerated at a number of sites, and it may represent a better alternative to the diazirine-containing amino acid.

Finally, we incorporated "caged-cysteine" at position α70 of the nAChR. In several experiments, caged-cysteine produced whole-cell currents of 1–2 μA in response to 200 μM ACh. Caged-cysteine is attractive because it allows one to block all endogenous cysteines by reacting them with a maleimide-based reagent. Subsequent decaging of the cysteine side chain with 320 nm light will reveal a unique cysteine, which may react with a wide variety of commercially-available maleimide or methanethiosulfonate (MTS) reagents. Caged cysteine may eventually provide a complementary method for site-specific cell-labeling, which may be particularly useful for labeling cells with multiple fluorescent probes for fluorescence-resonance energy transfer experiments [24].

Bibliography

[1] J.S. Debenham, R. Madsen, C. Roberts, and B. Fraser-Reid. 2 new orthogonal amine-protecting groups that can be cleaved under mild or neutral conditions. *J. Am. Chem. Soc.*, 117(11):3302–3303, 1995.

[2] L. Zervas, D. Borovas, and E. Gazis. *J. Am. Chem. Soc.*, 85:3660, 1963.

[3] B. Amit, U. Zehavi, and A. Patchornik. Photosensitive protecting groups of amino sugars and their use in glycoside synthesis. 2-Nitrobenzyloxycarbonylamino and 6-nitroveratryloxycarbonylamino derivatives. *J. Org. Chem.*, 39:192–196, 1974.

[4] U. Ellervik and G. Magnusson. 2-(Trimethylsilyl)ethyl glycosides. 9. Transformation into the corresponding 1-O-acyl sugars. *Acta. Chem. Scand.*, 47(8):826–828, 1993.

[5] C. J. Noren, S. J. Anthony-Cahill, M.C. Griffith, and P.G. Schultz. A general method for site-specific incorporation of unnatural amino acids into proteins. *Science*, 244:182–188, 1989.

[6] S. A. Robertson, J. A. Ellman, and P. G. Schultz. A general and efficient route for chemical aminoacylation of transfer RNAs. *J. Am. Chem. Soc.*, 113:2722–2729, 1991.

[7] J. Ellman, D. Mendel, S. Anthony-Cahill, C. J. Noren, and P. G. Schultz. Biosynthetic method for introducing unnatural amino-acids site-specifically into proteins. *Meth. Enzym.*, 202:301–336, 1991.

[8] M. Nassal. 4'-(1-azi-2,2,2-trifluoroethyl)phenylalanine, a photolabile carbene-generating analog of phenylalanine. *J. Am. Chem. Soc.*, 106(24):7540–7545, 1984.

[9] L. B. Shih and H. Bayley. A carbene-yielding amino acid for incorporation into peptide photoaffinity reagents. *Anal. Biochem.*, 144:132–41, 1985.

[10] M. Platz, A. S. Admasu, S. Kwiatkowski, P. J. Crocker, N. Imai, and D. S. Watt. Photolysis of 3-aryl-3-(trifluoromethyl)diazirines: A caveat regarding their use in photoaffinity probes. *Bioconjug. Chem.*, 2:337–41, 1991.

[11] S. K. Silverman. *I. Conformational and Charge Effects on High-Spin Organic Polyradicals; II. Studies on the Chemical-Scale Origin of Ion Selectivity in Potassium Channels.* Ph.D. Thesis, California Institute of Technology, 1998.

[12] M.C. Pirrung and D.S. Nunn. Synthesis of photodeprotectable serine derivatives – caged serine. *Bioorg. and Med. Chem. Lett.*, 2(12):1489–1492, 1992.

[13] M. W. Nowak, J. P. Gallivan, S. K. Silverman, C. G. Labarca, D. A. Dougherty, and H. A. Lester. *In vivo* incorporation of unnatural amino acids into ion channels in a *Xenopus* oocyte expression system. *Meth. Enzymol.*, 293:504–529, 1998.

[14] A. P. West Jr. *Theoretical Studies of Molecular Magnetism and Molecular Recognition, and Experimental Studies of the Extracellular Domain of the Nicotinic Receptor.* Ph.D. Thesis, California Institute of Technology, 1998.

[15] K.W. Rickert and B. Imperiali. Analysis of the conserved glycosylation site in the nicotinic acetylcholine receptor – potential roles in complex assembly. *Chemistry & Biology*, 2(11):751–759, 1995.

[16] B. Imperiali and K.W. Rickert. Conformational implications of asparagine-linked glycosylation. *Proc. Natl. Acad. Sci., U.S.A.*, 92(1):97–101, 1995.

[17] M. Quick and H. A. Lester. *Methods for Expression of Excitablitity Proteins in Xenopus Oocytes.* Academic Press, San Diego, CA, 1994.

[18] D.M. Lynn, S. Kanaoka, and R.H. Grubbs. Living ring-opening metathesis polymerization in aqueous media catalyzed by well-defined ruthenium carbene complexes. *J. Am. Chem. Soc.*, 118(4):784–790, 1996.

[19] D.M. Lynn, B. Mohr, and R.H. Grubbs. Living ring-opening metathesis polymerization in water. *J. Am. Chem. Soc.*, 120(7):1627–1628, 1998.

[20] V.W. Cornish, K.M. Hahn, and P.G. Schultz. Site-specific protein modification using a ketone handle. *J. Am. Chem. Soc.*, 118(34):8150–8151, 1996.

[21] W. Zhong. *Physical Organic Chemistry on the Nicotinic Acetylcholine Receptor.* Ph.D. Thesis, California Institute of Technology, 1998.

[22] L.K. Mahal, K.J. Yarema, and C.R. Bertozzi. Engineering chemical reactivity on cell surfaces through oligosaccharide biosynthesis. *Science*, 276(5315):1125–1128, 1997.

[23] V.W. Cornish, D.R. Benson, C.A. Altenbach, K. Hideg, W.L. Hubbell, and P.G. Schultz. Site-specific incorporation of biophysical probes into proteins. *Proc. Natl. Acad. Sci., U.S.A.*, 91(8):2910–2914, 1994.

[24] L.M. Mannuzzu, M.M. Moronne, and E.Y. Isacoff. Direct physical measure of conformational rearrangement underlying potassium channel gating. *Science*, 271(5246):213–216, 1996.